American Psychological Association

Volume 109 Number 7 October 2017

Published eight times ISSN 0022-0663

Journal of Educational Psychology

www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/pubs/journals/subscriptions.aspx

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/pubs/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Steve Graham, at steve.graham@asu.edu. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

Electronic Access: APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES® full-text database. See http://my.apa.org/access.html.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

APA Journal Staff: Rosemarie Sokol-Chang, PhD, *Publisher, APA Journals*; Mare Meadows, *Managing Director*; Amanda S. Conley, *Journal Production Manager*; Cheryl Johnson, *Editorial Manuscript Coordinator*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits.*

Journal of Educational Psychology® (ISSN 0022-0663) is published eight times (January, February, April, May, July, August, October, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2017 rates follow: *Nonmember Individual:* $250 Domestic, $292 Foreign, $314 Air Mail. *Institutional:* $953 Domestic, $1,030 Foreign, $1,054 Air Mail. *APA Member:* $123. *APA Student Affiliate:* $75. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

Journal of

Educational Psychology®

www.apa.org/pubs/journals/edu

October 2017

Volume 109
Number 7

Reading and Writing

Mathematics

Teachers Social and Emotional Competence

(*Contents continue*)

Journal of Educational Psychology
2017, Vol. 109, No. 7, 889–914

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000181

Early Intervention for Children at Risk for Reading Disabilities: The Impact of Grade at Intervention and Individual Differences on Intervention Outcomes

Maureen W. Lovett
The Hospital for Sick Children, Toronto, Ontario, Canada, and University of Toronto

Jan C. Frijters
Brock University

Maryanne Wolf
Tufts University

Karen A. Steinbach
The Hospital for Sick Children, Toronto, Ontario, Canada

Rose A. Sevcik and Robin D. Morris
Georgia State University

Across multiple schools and sites, the impact of grade-at-intervention was evaluated for children at risk or meeting criteria for reading disabilities. A multiple-component reading intervention with demonstrated efficacy was offered to small groups of children in 1st, 2nd, or 3rd grade. In a quasi-experimental design, 172 children received the Triple-Focus Program (PHAST + RAVE-O), and 47 were control participants. Change during intervention and 1–3 years later (6–8 testing points), and the influence of individual differences in predicting outcomes, were assessed using reading and reading-related repeated measures. Intervention children out-performed control children at posttest on all 14 outcomes, with average effect sizes (Cohen's *d*) on standardized measures of .80 and on experimental measures of 1.69. On foundational word reading skills (standardized measures), children who received intervention earlier, in 1st and 2nd grade, made gains relative to controls almost twice that of children receiving intervention in 3rd grade. At follow-up, the advantage of 1st grade intervention was even clearer: First graders continued to grow at faster rates over the follow-up years than 2nd graders on 6 of 8 reading outcomes. For some outcomes with metalinguistic demands beyond the phonological, however, a posttest advantage was revealed for 2nd Grade Triple participants and for 3rd Grade Triple participants relative to controls. Estimated IQ predicted growth during intervention on 7 of 8 outcomes. Growth during follow-up was

This article was published Online First March 23, 2017.

Maureen W. Lovett, The Hospital for Sick Children, Toronto, Ontario, Canada, and Department of Paediatrics, University of Toronto; Jan C. Frijters, Department of Child and Youth Studies, Brock University; Maryanne Wolf, Eliot-Pearson Department of Child Study and Human Development, Tufts University; Karen A. Steinbach, The Hospital for Sick Children, Toronto, Ontario, Canada; Rose A. Sevcik and Robin D. Morris, Department of Psychology, Georgia State University.

In Atlanta, Boston, and Toronto, we are especially grateful to the 237 children who participated in this research for their interest, enthusiasm, and effort. We also gratefully acknowledge their parents, teachers, and schools for their commitment and support during the course of this study. The cooperation and contribution of the principals and staff of all the participating schools, all of whom offered space and opportunities for our programs, are much appreciated.

In Atlanta, we acknowledge the students and their families from the Fulton County School system, administrators, participating schools, their principals and staffs, particularly Susan Grabel; our research teachers, Eileen Cohen, Mary Bucklen, Kim Imbrecht, Victoria Burke, Heather Lubeck, Cashawn Myers, Judith Mahoney, and Nioyonu Olutosin; and members of our research team, Paul Cirino, Marla Shapiro, Cynthia Martin, Nicole Mickley, Becky Doyle, Justin Wise, Hye K. Pae, Jennifer Harrison, and Laina Jones.

In Boston, we acknowledge the students and their families from Somerville, Medford, and Newton School systems, administrators, particularly Alice O'Rourke and Roy Belson, participating schools, their principals and staffs; our research teachers, Katharine Donnelly-Adams, Terry Joffe Benaryeh, Joanna Christodoulou, Fran Lunney, Anne Knight, Jill Ludmar, Jane Hill-Lovins, and Andrea Marquant; and members of our research team, Beth O'Brien, Cathy Moritz, Julie Jeffery, Lynne Miller, Alyssa Goldberg O'Rourke, Chip Gidney, Wendy Galante, Tami Katzir, Alexis Berry, Laura Vanderberg, Ellen Boiselle, Sasha Yampolsky, and Gordon Goodman.

In Toronto, we acknowledge the children and their families from the Toronto District School Board, the Toronto Catholic District School Board, and the Peel District School Board; our Senior Teacher Trainer and Program Developer, Léa Lacerenza; our research teachers, Denis Murphy, Jody Chong, Tammy Cohen, Vicky Grondin, and Steacy O'Connor; and our psychology assessment team, Jennifer Janes, Jennifer Goudey, Leslie Daniels, Jennifer McTaggart, and Jennifer Lasenby; and our senior research team members, Maria De Palma, Meredith Temple, and the late Nancy Benson.

The research reported here was supported by a National Institute of Child Health and Human Development Grant HD30970 to Georgia State University, Tufts University, and The Hospital for Sick Children/University of Toronto.

Correspondence concerning this article should be addressed to Maureen W. Lovett, Neurosciences and Mental Health, Director, Learning Disabilities Research Program, The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario, Canada M5G 1X8. E-mail: mwl@sickkids.ca

predicted by vocabulary and visual sequential memory. These findings provide evidence on the importance of early intensive evidence-based intervention for reading problems in the primary grades.

Educational Impact and Implications Statement
Does it matter in what grade early reading intervention is provided for young children who are struggling with learning to read—in 1st, or 2nd, or 3rd grade? Children with reading disabilities (RD), or at risk of RD, were taught in small groups an hour a day for 125 hours using a reading intervention developed and found effective in our earlier research; the children received this program either in 1st or 2nd or 3rd grade. All children improved their reading after receiving this program when compared to other children with RD who received whatever their schools offered. Children who received the program in 1st or 2nd grade made greater gains in basic reading skills than those who received it in 3rd grade; and those who received it in 1st grade continued to develop reading at faster rates well after the program ended. These findings provide evidence for the importance of early intensive reading intervention for struggling readers, and support intervention starting in 1st grade.

Keywords: reading, reading disabilities, early intervention, outcomes, follow-up

Several landmark studies reported in the late 1990s compared different approaches to the remediation and/or prevention of reading acquisition problems in the early elementary grades (Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Foorman et al., 1997; Scanlon & Vellutino, 1997; Torgesen, Wagner, & Rashotte, 1997a; Torgesen et al., 1999; Vellutino et al., 1996). Research by Foorman and her colleagues (1998) provided important evidence that explicit classroom instruction in letter-sound correspondences can prevent reading failure in 1st and 2nd grade children at risk for reading problems. Another classroom-based intervention, Peer-Assisted Learning Strategies (PALS), developed by Doug and Lynn Fuchs, yielded positive results on measures of word recognition and text reading skill, and was found to improve reading skills for both struggling and average readers (Fuchs & Fuchs, 2005; Mathes, Howard, Allen, & Fuchs, 1998). These results demonstrate that opportunities exist to provide targeted and differentiated instruction within the classroom setting to reduce the prevalence of reading problems (Fletcher, Lyon, Fuchs, & Barnes, 2007; Fuchs, Fuchs, Mathes, & Simmons, 1997).

Torgesen, Wagner, Rashotte, et al. (1999) also reported seminal early intervention research, but their work involved one-on-one remedial intervention outside of the classroom over a period of 2 1/2 years. Their participants were at-risk children, those lowest in letter naming and phonological awareness, entering kindergarten. At the end of the 2 1/2 year intervention, children who had received explicit phonological awareness and synthetic phonics training were the strongest readers when average group scores were assessed. Both their nonword reading (Word Attack) and word reading (Word Identification) skills fell overall within the average range. Their advantage relative to other groups was not consistently established, however, on all dimensions of reading skill at the end of second grade, suggesting that early intervention efforts may not yield equivalent impact on different reading-related processes (Torgesen et al., 1999, 2001; Torgesen, Wagner, Rashotte, Alexander, & Conway, 1997).

Overall, these landmark studies and those that have followed have provided strong converging evidence for the efficacy and cost-effectiveness of early intervention efforts (Al Otaiba, 2000; Berninger et al., 2000, 2002; Connor et al., 2007; Mathes et al., 2005; O'Connor, 2000; O'Connor, Fulmer, Harty, & Bell, 2005; Vaughn, Linan-Thompson, & Hickman, 2003). Foorman and Al Otaiba (2009) contend that better classroom instruction can reduce the number of low-achieving children to around 5%, and further supplemental small group or individual tutoring can bring the numbers down even lower to 1%–3%. Evaluations of multitiered intervention models have suggested rates of inadequate response could be as low as 2%–5% with effective and well-timed early reading intervention (Berninger et al., 2003; Mathes et al., 2005; McMaster, Fuchs, Fuchs, & Compton, 2005; Torgesen, 2000).

Such efforts require coordinated infrastructure and investment at the school level, including universal early screening for academic risk, access to effective early intervention within the school, teacher preparation and confidence in such procedures, regular progress monitoring for all children, and access to booster interventions when needed (Fletcher & Vaughn, 2009). The growing literature on response to intervention (RTI) provides guidance on how this infrastructure can be put in place and its benefits, while also recognizing the challenges still to be addressed (Denton, Fletcher, Anthony, & Francis, 2006; Fletcher & Vaughn, 2009; Fuchs & Fuchs, 1998; Glover & Vaughn, 2010; Vaughn & Fuchs, 2003; Vaughn et al., 2011).

How Important Is the Timing of Early Intervention?

Despite evidence of the effectiveness of early intervention for children at risk for reading failure, relatively few empirical studies have been conducted to compare the relative efficacy of reading intervention initiated at different ages. The majority of early intervention research has been conducted with at-risk children in kindergarten or first grade and has reported positive outcomes. A meta-analysis by Wanzek and Vaughn (2007) summarized evidence from early intervention studies offering at least 100 sessions. There were diverse outcomes reported from this work, although the majority described effect sizes in the moderate-to-large range. In this review, effect sizes were found to be larger for intervention studies conducted with Kindergarten and first graders (average effect size ranging from .31 to .84) than with children in 2nd and 3rd grades (average e.s. .23–.27). Scammacca et al. (2007) in their report suggested that gains from early interventions of longer duration tend to be maintained at least until the 2nd grade. Vadasy

and colleagues provided two years of intervention in 1st and 2nd grade, and reported average effect sizes of .64 on reading outcomes, suggesting that longer duration of intervention was also relevant to benefits of early intervention (Vadasy, Sanders, Peyton, & Jenkins, 2002). It is noted that the meta-analysis cannot address causal evidence examining the effects of duration, intensity, or timing of intervention and that the studies assessed varied greatly in the extent to which interventions were operationalized.

Evidence From Classroom Intervention Studies

The only experimentally controlled study to date of the timing of reading intervention was reported by Connor and her colleagues (Connor, Morrison, Fishman, Crowe, et al. 2013). These researchers asked whether the timing and duration of individualized reading instruction within the classroom would make a difference to children's reading achievement by the end of 3rd grade. The reading intervention included teacher professional development and instruction individualized through computer software to match the student's performance on word reading, vocabulary, and comprehension assessments; this intervention has been found to yield positive effects in smaller efficacy trials conducted within single grade levels (A2i—Connor et al., 2007; Connor, Morrison, Schatschneider, et al., 2011; Connor, Morrison, Fishman, et al., 2011). The intervention itself varies the amount of instructional time allocated different instructional components and types of reading activities. Although not a supplemental reading intervention, Individualizing Student Instruction (ISI) individualizes instruction by differentially weighting instructional time and components within the classroom.

In Connor et al.'s (2013) study, the influence of ISI was evaluated longitudinally over three grades in a cluster-randomized controlled design; classrooms were randomly assigned to ISI treatment or control conditions, and teachers in the control condition received an equivalent amount of professional development and attention in a mathematics intervention condition. Randomization of classrooms to condition occurred every year, so there were children who received one, two, or three years of ISI, with ISI beginning in their 1st, 2nd, or 3rd grade year. Both the timing and duration of the ISI intervention therefore could be evaluated. Results revealed an overall advantage for those children who received ISI reading instruction throughout Grades 1–3; these children on average were performing above grade level by the end of 3rd grade, demonstrating the advantage of accumulated benefit (e.s. = .90 relative to three years control placement). There was an advantage for 1st grade intervention: Those who received only one year of ISI in 1st grade outperformed those whose year of ISI occurred in 2nd or 3rd grade. Connor and colleagues note, however, that the first grade advantage was "inconsistent" and was not replicated for children who received two years of ISI. For these children, there was greater benefit to receiving ISI in Grades 1 and 3 rather than Grades 1 and 2 or Grades 2 and 3.

Rather than a supplementary or pull-out remedial intervention, Individualizing Student Instruction (ISI) is a weighting of instructional time and components within the general classroom (Connor et al., 2007, 2013; Connor, Morrison, et al., 2011). Many early intervention studies are classroom studies and typically recruit entire classes as samples. The emphasis is one of preventing reading failure through early intense instruction for at-risk children. These studies do not focus on samples of struggling learners and may be expected to produce different patterns of findings than those that recruit samples performing substantially below age level expectations.

Evidence From Supplemental Intervention Studies

In contrast, other early intervention studies have involved well-defined supplemental programs for at-risk learners and have operationalized these interventions. It should be noted that almost all of this research has avoided evaluation of school-based special education programs. In fact, as Fletcher and Vaughn (2009) point out, outcome data from evaluations of at-risk learners in special education placements have not been encouraging—many reports have documented limited growth and poor outcomes, suggesting typical interventions in many special educational settings to be generally ineffective in terms of accelerating academic growth (Hanushek, Kain, & Rivkin, 1998; Morgan, Frisco, Farkas, & Hibel, 2010; Vaughn, Levy, Coleman, & Bos, 2002). As Fletcher and Vaughn suggest, "There is a major disconnection between what is known about efficacy of instruction for students with academic difficulties and how students are taught in schools, especially for students most at risk for academic and behavioral difficulties" (p. 33).

A recent report using a national U.S. dataset (the Early Childhood Longitudinal Study-Kindergarten Cohort) provides quite a different perspective on the potential effectiveness of special education placement. Ehrhardt, Huntington, Molino, and Barbaresi (2013) were interested in determining whether grade at entry to special education was related to reading growth between 1st and 5th grades in a sample of children with reading problems. Lacking standard measurement for reading disorders, these investigators selected children from the cohort who had an IEP targeting reading and for whom the special education teacher listed specific learning disability as the primary category of disability. Early entry to special education proved significantly associated with reading achievement: Children entering before or during 1st grade demonstrated superior reading achievement gains to those who entered in 2nd or in 3rd grade. Of interest, these early entry children were not significantly different from those who entered special education in 4th or 5th grade however; it is acknowledged that the latter students may have had less severe reading impairment than those with earlier identification (Leach, Scarborough, & Rescorla, 2003).

Another meta-analysis was undertaken with a developmental perspective, focusing on the interaction of grade and intervention modality to assess moderators of intervention efficacy (effect sizes) for at-risk and struggling readers. Suggate (2010) was interested in whether intervention effect sizes varied with grade at intervention (preschool through Grade 7) and type of intervention offered (phonics, comprehension, or mixed focus). Overall he reported that reading intervention was associated with clear improvement in reading outcomes following intervention both in the immediate short-term ($d = 0.49$) and over the longer-term ($d = 0.36$). Overall effect sizes were found to be greater for older children (Gr 5–7: $d = 0.68$) than for children in preschool and kindergarten ($d = 0.36$), Grades 1 and 2 ($ds = 0.52$ and 0.54), and Grades 3 and 4 ($d = 0.59$). Mixed and comprehension-focused interventions were associated with greater effect sizes for older children, and phonics interventions for children in kindergarten and 1st grade.

Suggate's finding of greater effects with older readers stands in contrast to that of most previous studies and should be considered

in light of how effect sizes are calculated in his report. Mean effect sizes were calculated for each outcome measure from the original 85 studies included in the meta-analysis: intervention group performance minus control group performance divided by the pooled standard deviation of the two groups (Cohen's *d*; Hunter & Schmidt, 2004). Lower effect sizes for children in the younger grades may indicate not that they did not gain with intervention but that control participants in those grades also made reading gains without the intervention. Such an interpretation is possible given that Suggate (2010) describes significant negative correlations between grade and control group standard scores, illustrating that older children were more impaired relative to norms.

Measurement Issues Complicate the Interpretation of Intervention Effects for Different Reading Outcomes

Evaluating timing-of-intervention effects is made much more complex by the interaction between age at intervention and the type of reading outcome being evaluated, as well as differences in the ability to reliably *measure* different dimensions of reading skill at different ages. There is ample evidence of reading interventions having demonstrated efficacy on some, but not all, dimensions of reading skill. Meta-analyses such as those conducted by Scammacca, Edmonds, and their colleagues have revealed marked variability in reading comprehension remediation effects across studies, and that at least with older students, average gains in reading comprehension with intervention were typically smaller than those seen on basic reading skills (Edmonds, Vaughn, Wexler, et al., 2009; Scammacca et al., 2007). Connor's data on the timing of her ISI intervention with younger readers (Grades 1, 2, 3) also demonstrated some variability in effect size estimates for word identification versus comprehension outcomes in 3rd grade (Connor et al., 2013). For 1st and 2nd graders, effect sizes for ISI instruction were roughly equivalent for word identification and passage comprehension outcomes (1st grade Cohen's d = .32 and .36; 2nd grade d = .44 and .43, respectively), whereas for 3rd graders, ISI yielded effect sizes of .26 on word identification, but only .06 on passage comprehension. Similarly the Reading First Impact Report noted that although the Reading First initiative had no effect on the reading comprehension scores of students in Grades 1, 2, or 3, small positive effects on decoding skills were observed for the subsample of 1st graders studied (Gamse, Jacob, Horst, Boulay, & Unlu, 2008). Another higher-level reading outcome has also demonstrated variable outcomes following early intense reading intervention—reading rate or fluency. When Torgesen collected follow-up data on his intervention participants at 8 and 10 years of age, he found that they exhibited substantial deficits in reading rate despite otherwise positive reading outcomes (Torgesen et al., 2001).

Part of the problem concerns the continuing difficulty in developing appropriate measurement for more complex dimensions of reading skill like text comprehension and reading fluency. Outcome measures that have been used across studies have varied enormously in their power and sensitivity to intervention-related change. Many investigators have acknowledged that reading interventions typically yield larger effects on researcher-developed than standardized measures (Edmonds et al., 2009; Lovett, Barron, & Frijters, 2013; Swanson, Hoskyn, & Lee, 1999). Experimental measures with more trials per level of difficulty result in more visible gains and better opportunities to demonstrate intervention-related change over the short-term. Many questions remain regarding the best measurement models for evaluating an intervention that targets functions as complex as reading comprehension and fluency.

Individual Variability in Response to Early Intervention

Another contributor to variability in response to early intervention may stem from individual differences among the children receiving reading intervention. There have been some efforts to determine whether children with different cognitive and academic profiles respond differently to early intervention and some evidence to support that suggestion (Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; Frijters et al., 2011; Foorman et al., 1998; Nelson, Benner, & Gonzalez, 2005). Al Otaiba and Fuchs (2002) reviewed 21 studies examining nonresponders to reading intervention; they found that although seven child characteristics had been related to nonresponse, phonological awareness was the most consistently correlated across studies. These investigators subsequently conducted a longitudinal study with kindergarteners and 1st graders and found that a combination of naming speed, vocabulary, sentence imitation, problem behavior, and amount of reading intervention correctly predicted 82% of nonresponsive students and 84% of always responsive students (Al Otaiba & Fuchs, 2006). Inconsistent responders were predicted far less reliably (30%), focusing attention on the issue of how treatment response is actually defined and operationalized, an issue Frijters and our group recently assessed with older disabled readers (Frijters, Lovett, Sevcik, & Morris, 2013).

Multiple Component Interventions in the Remediation of Reading Disabilities

In addition to measurement issues, many intervention reports over the past two decades have suggested that progress in remediating phonological decoding deficits has not been matched by gains in fluency and reading comprehension. Compton, Miller, Elleman, and Steacy (2014) noted recently that successfully learning how to decode a new word does not ensure that that word will come to be integrated into what has been called "a rich orthographic reading vocabulary" (Torgesen, Wagner, & Rashotte, 1997b). Torgesen suggested that the problem reflects the complexity of the processing impairments seen in more severely disabled readers (Torgesen et al., 1997b). Such children require intervention that includes systematic and explicit phonological decoding instruction, but also offers focused remedial components to address other areas of deficit.

Researchers have demonstrated that many children with RD experience particular difficulties with strategy learning and the acquisition of self regulatory strategies, and that these problems appear to exist independent of their phonological difficulties (Swanson & Sáez, 2003; Swanson, Sáez, & Gerber, 2006; Swanson & Siegel, 2001). The strategy deficits of children at risk for reading acquisition failure extend beyond the word identification and word attack foundations of literacy and encompass all aspects of reading for meaning, expository text comprehension, and written expression. There is evidence that low-achieving readers can make gains in reading comprehension with systematic instruction and practice on specific reading comprehension strategies (Mason,

2004; Vaughn et al., 2000). It is reasonable to hypothesize that explicit strategy training and metacognitive instruction could be used to address and prevent generalization failures, and provide an important component of effective remediation for RD in the acquisition of both decoding and reading comprehension skills.

Lovett, Lacerenza, Borden, et al. (2000) reported evidence supporting this speculation: When a phonological reading intervention (PHAB/DI for Phonological Analysis and Blending) was combined with the teaching of specific word identification strategies, and these strategies were implemented, practiced, and evaluated using self-directing dialogue (WIST for Word Identification Strategy Training), severely disabled readers demonstrated superior reading achievement and faster learning than when they received an equal amount of intervention in phonological or strategy training conditions separately. The combined intervention conditions were associated with the greatest generalization of gains for these children with severe RD. These results provided evidence of the importance of strategy instruction to effective remediation and led to integration of these two interventions into the PHAST (Phonological and Strategy Training) Reading Program (Lovett, Lacerenza, & Borden, 2000). In earlier work, the authors had demonstrated in a controlled evaluation the efficacy of both the PHAB and WIST Programs relative to a control program, and some program-specific effects (Lovett et al., 1994).

The need for a multidimensional perspective on the core processing deficits of children with RD is echoed in the work of Wolf and her colleagues who identify naming speed deficits as a window on the failure of struggling readers to build integrated, rapid, and automatic connections among the component processes necessary to fluent reading acquisition (Wolf & Bowers, 1999; Wolf & Katzir-Cohen, 2001). Citing the support of research on the importance of high quality orthographic, semantic, morphosyntactic, and phonological lexical representations, and of their interconnections, Wolf and colleagues developed a reading intervention designed to strengthen lexical representational systems and teach explicitly the connections among representations. Called RAVE-O (for Retrieval, Automaticity, Vocabulary, Engagement with Language, and Orthography), the program seeks to teach young readers to enrich and connect all their knowledge about a word as quickly as possible. The idea is to simulate what typically developing brain circuitry does during the early stages of reading development (Wolf et al., 2009). RAVE-O is designed to accompany a systematic program of phonologically-based decoding instruction, and is directed to development of an appreciation of the richness of oral and printed language, and an enjoyment of words and reading for meaning.

Both the PHAST and the RAVE-O Reading Programs have been evaluated in a previous multisite intervention study conducted by the present authors (Morris et al., 2012). This previous study included 279 2nd and 3rd grade children meeting low achievement or IQ-reading discrepancy definitions of RD (the majority meeting both criteria), and with diverse demographic profiles (IQ, SES, race). Children were randomly assigned to program according to a $2 \times 2 \times 2$ factorial design according to the demographic variables of IQ (70–89; 90+), SES (low; average), and race (Black; Caucasian). The effectiveness of two multiple-component intervention programs for children with RD (PHAB/DI + RAVE-O (Wolf et al., 2000); and the PHAST Reading Program (Lovett, Lacerenza, & Borden, 2000) were evaluated against both an alternative treatment control program (Classroom Survival Skills (CSS) + Math), and a phonological treatment program paired with CSS (PHAB/DI + CSS). Interventions were taught an hour daily for 70 days on a 1:4 ratio at 3 different sites (Atlanta, Boston, Toronto).

Results indicated that both the PHAST and the RAVE-O (+ PHAB/DI) Programs were associated with significant improvement on basic reading skills relative to the alternative control group and the phonological treatment group at the end of the program and at 1-year follow-up testing (Morris et al., 2012). Equivalent gains were observed for children of different racial, SES, and IQ groups; these factors did not systematically interact with treatment program and did not differentially predict outcomes at either posttest or at 1-year follow-up. Both multiple-component programs were confirmed to be effective vehicles of intervention for struggling readers from a wide range of backgrounds and with differing levels of intellectual functioning. Differential treatment outcome effects were found between the multidimensional programs at posttesting based on the respective emphases of the programs.

In the present study, the PHAST and RAVE-O Programs were integrated to produce what is called the Triple-Focus Program, designed to capitalize upon the positive effects associated with both multiple component programs. (The PHAST Program is considered a "double program" because it integrates the PHAB and WIST Programs.) The Triple-Focus Program provides tailored and intensive remediation that combines explicit phonological instruction with word identification strategy training, reading comprehension strategy training, and instructional activities that foster enriched lexical representations and increased engagement with word play, reading, and text comprehension.

Questions Motivating the Present Study

The present study was undertaken to evaluate issues related to the timing of reading intervention for children meeting criteria for reading disabilities at the end of 1st or 2nd grade, or meeting risk criteria at the end of kindergarten. A full year of small group intervention using the Triple-Focus Program was provided for a total of 100–125 instructional hours in Grades 1, 2, or 3. The questions addressed in the present study included:

1. Did grade at intervention influence treatment outcomes and rate of growth in the short-term and/or over follow-up?
2. Did grade at intervention influence rate of normalization of reading scores following intervention?
3. Were there individual differences in cognitive and reading-related profiles that influenced response to intervention in the short- and long-term?

Method

Study Design

This present design evaluated the impact of developmental timing of reading intervention (1st, 2nd, or 3rd grade), longitudinal

change in a repeated measurement design (testing at 0, 35, 70, 105, 125 hr of instruction, and at 1–3 years follow-up), and the role of individual differences on short- and long-term reading outcomes. The experimental reading intervention used here integrates two research-based remedial reading programs with demonstrated efficacy (PHAST + RAVE-O) into a comprehensive Triple-Focus intervention. The Triple-Focus intervention employed the same format as our previously reported interventions PHAST and RAVE-O (Morris et al., 2012), was taught by trained research teachers hired for the project, and was independently monitored for treatment integrity. The program was a pull-out intervention taught on a 1:4 ratio for an hour a day in the child's home school. Reading outcomes for the Triple-Focus participants were compared with those of curricular or business-as-usual controls in the same grades and with the same degree of reading and reading-related impairment. Because participants were not consistently assigned randomly to intervention or control condition, this is considered a quasi-experimental design.

Participants. Participants were recruited from multiple schools in three large metropolitan areas (Atlanta, Boston, and Toronto) on the basis of teacher referral for significant underachievement in reading. General inclusion criteria consisted of: English as their first and primary language, enrolment in 1st, 2nd, or 3rd grade at time of teacher referral, and normal or corrected hearing and vision. A total of 416 children with reading problems were referred for screening across the three study sites to see if they would qualify for participation.

All participants were required to meet specified exclusion and inclusion criteria. Children who had histories of hearing impairment (>25dB at 500 + Hz bilaterally), of uncorrected visual impairment (>20/40), serious emotional/psychiatric disturbance (i.e., psychotic, pervasive developmental disorder), or chronic medical/neurological conditions (i.e., uncontrolled seizure disorder, congenital heart disease, acquired brain injuries) were excluded based on a brief demographic and history form completed by their parents. In addition, children were excluded if they had repeated a grade or received a K-BIT composite score below 70. The repetition of a grade was an exclusionary criterion because of our attempt to recruit at each grade level groupings of the same age; in practice, grade retention was very rarely seen in participating schools. The co-occurrence of ADHD, a disorder common in RD populations, did not exclude a child from participation.

Participants were further selected from this pool based on their performance on a screening battery that included the Kaufman Brief Intelligence Test (K-BIT; Vocabulary and Matrices; Kaufman & Kaufman, 1990), Woodcock Reading Mastery Test—Revised (WRMT-R; Woodcock, 1987), and the Wide Range Achievement Test-3rd Edition–Reading (WRAT-3; Wilkinson, 1993). Subtests from the WRMT-R included Word Identification, Word Attack, and Passage Comprehension; for those children screened at the end of Kindergarten or beginning of Grade 1, WRMT-R Visual-Auditory Learning and Letter Identification were also administered although almost no children qualified solely on the Readiness Cluster score to which these subtests contribute.

All children selected for participation qualified based on meeting a low-achievement criterion for reading disabilities; this criterion required reading performance at or below a composite standard score of 85 on multiple standardized reading measures. Reading performance was measured using one or more of the following indices: (a) a Reading Total score calculated by averaging the standard scores on the WRMT-R Passage Comprehension, Word Identification, Word Attack, and WRAT-3 Reading subtests; (b) the WRMT-R Basic Skills Cluster score; (c) and/or the WRMT-R Total Reading Cluster score (Short Scale). The Basic Skills Cluster Score is the composite of Word Identification and Word Attack; the Total Reading Cluster–Short Form is the composite of Word Identification and Passage Comprehension. Children qualified by demonstrating low achievement on one of these three reading indices—that is, at least one of their reading composite standard scores was 85 or below (at or less than the 16th %tile). Of the participants who qualified for inclusion, 63% met all three criteria, 18% met two, and 19% met one.

Children of any race or ethnic group, or either sex, were included as long as they met the English as the primary language requirement and the low achievement criterion for RD. We sought to include diverse samples of children, with the goal of including large numbers of minority children, girls, and children from low SES families. Given that our studies were located within public schools in three major cities, obtaining this level of minority children involvement was not difficult, although obtaining samples with 50% girls proved to be more difficult (Morris et al., 2012).

SES was assessed by parental occupation and educational status using an index of the families' SES. SES data from all sites was derived using two American SES scales (Entwisle & Astone, 1994; Hollingshead, 1975; Nakao & Treas, 1992) and one Canadian SES scale (Blishen, Carroll, & Moore, 1987). Our goal was to develop an index for systematically identifying the children's families as average or above SES, or below average SES. The particular differences or actual levels of SES provided by the scales were not as critical as an accurate ranking of children. A systematic evaluation of the reliability and concordance of these different scales was undertaken and results were used to classify the children into the Average or Low SES groups based on a systematic combination of the different indices. Details of this work on SES measurement have been published (Cirino et al., 2002).

Of the 237 participants who met all criteria and were selected, 172 children participated in instructional groups in the Triple-Focus Program (79 Grade 1, 43 Grade 2, 51 Grade 3) and 47 served as control participants (18 Grade 1, 13 Grade 2, 16 Grade 3). Attrition was fairly low in the study given the length of the participation period (31 out of 237 enrolled cases, or a rate of 13%). A total of 17 children were lost to attrition between enrollment and the start of intervention, and an additional 14 were lost between pretest and posttest. Attrition generally was due to families relocating, switching schools, or having difficulties transporting children to the classes. Random assignment of children to intervention condition was not possible during the first and last years of data collection, and the design therefore should be considered quasi-experimental.[1] Children meeting criteria within a school were grouped together on the basis of grade and raw reading scores (WRMT-R Word Identification and Word Attack); the group was then proposed as an instructional group to the main

[1] In Year 1, a decision was made to start as many intervention classes as possible in the two sites developing content for programming (Toronto, Boston).

site in Atlanta. If accepted, the instructional group was assigned to the Triple-Focus intervention in the present study. The control group included participants who met all criteria for inclusion but failed to match into an instructional group: Any participant meeting criteria who did not match into an instructional group, or who was referred and screened after classes had started, or was from a school where other participants were not available to form an intervention class, was assigned to the control condition. Despite efforts to enroll more control participants, control numbers remained far lower than projected due to the difficulty in enrolling children with reading disabilities and having them wait a full year before they could access our intervention program. In the final year of data collection, there was a bias toward 'catching up' by attempting to add more control participants; this led to inclusion of some control participants from schools who did not have any Triple-Focus intervention classes running. The study ran over five school years in total. A flowchart provides an overview of recruitment, enrollment, assignment, and intervention for 1st, 2nd, and 3rd grade participants in Atlanta, Boston and Toronto (see Table A1 in the Appendix).

Table 1 displays overall reading scores and participant profiles for the sample. Results of a multivariate analysis of variance confirmed that intervention and control participants were comparable at pretest on all selection criteria, including age, $F(8, 199) = 1.44$, $p = .18$. Descriptive statistics for every outcome measure subdivided by time of test (pretest, posttest), intervention condition (Triple-Focus, Control), and grade (Grades 1, 2, 3) have been provided in Table A2 of the Appendix. Additional evidence of group comparability can be seen in Table 3, in which γ_{01} represents the test of intervention and control pretest differences on each outcome. The total sample was confirmed to be significantly impaired on all measures of reading achievement, performing more than one standard deviation below expectations on measures of decoding, word reading, and passage comprehension, but at the lower end of the average range on measures of receptive vocabulary and intellectual functioning. Overall the sample was almost a full standard deviation below expectation on the Freedom from Distractibility factor score from the WISC, suggesting that a high proportion of participants may have had attention difficulties. Approximately 51% of the sample was from low SES families, and 64% were males.

Measures

The reading and reading-related measures below were selected because they are standardized, widely used in educational and intervention outcome research, and psychometrically appropriate for growth-curve modeling. Use of these and the experimental measures allow for comparison with our own past research and other major intervention studies in the literature. The robust psychometric properties of the experimental measures of learning and transfer-of-learning have been documented in a separate report by our group (Cirino et al., 2002). As well, the standardized measures have been selected because of similar excellent psychometric characteristics, including reliability, construct validity, and the ability to sensitively measure change in reading and related skills. Measures were administered by Masters-level research assistants or senior graduate students trained and supervised by the Research Coordinator in each site. Examiners were only allowed to test independently after completing training, observing a trained examiner, and being observed by the Research Coordinator during testing. Double scoring was used to ensure the accuracy of scoring and to assess interscorer reliability.

Standardized measures of reading and related skills. The intervention children were tested at pre, mid, post, and follow-up; the controls were tested at pretest and posttest only.

Word reading. The first measure used is *WRMT-R, Form G* (Woodcock, 1987)—Word Identification subtest. The Word Identification subtest presents letters and then words in isolation for students to identify. *Wide Range Achievement Test-3 (WRAT-3).* The WRAT-3 (Wilkinson, 1993) similarly measures both individual letter identification and word reading (Reading). Test–retest reliability exceeds .95 for the WRMT-R Word Identification subtest; alternate form reliability exceeds .87 for the WRAT-3 Reading subtest.

Speeded word identification. The measure used is *Test of Word Reading Efficiency (TOWRE), Sight Word Efficiency* subtest (Torgesen, Wagner, & Rashotte, 1999). Sight Word Efficiency assesses the number of real words that can be accurately read within 45 seconds. Alternate-form reliability exceeds .88 for Sight Word Efficiency.

Nonword decoding. Measures used are *WRMT-R Word Attack* subtest and *TOWRE Phonemic Decoding Efficiency* subtest. On

Table 1
Participant Characteristics at Intervention Start (n = 219)

Characteristic	Intervention (n = 172) *M (SD)*	Control (n = 47) *M (SD)*
Age in months	89.2 (12.2)	91.5 (11.9)
WRMT-R Word Attack scaled score	76.6 (9.3)	74.2 (8.2)
WRMT-R Word Identification scaled score	81.1 (9.9)	77.7 (11.1)
WRMT-R Passage Comprehension scaled score	79.2 (8.9)	76.9 (10.6)
WRAT-III Reading scaled score	85.4 (10.0)	79.3 (10.0)
WISC Freedom from Distractibility Index	87.3 (12.6)	83.2 (12.6)
WISC Processing Speed Index	96.6 (15.2)	95.1 (13.8)
Kaufman Brief Intelligence Test	92.6 (10.4)	93.0 (10.6)
Peabody Picture Vocabulary Test	92.1 (13.6)	98.1 (14.9)
Proportion male	.624	.702
Number in Grade 1	79	18
Number in Grade 2	43	13
Number in Grade 3	51	16
Proportion low socioeconomic status	.487	.574

these measures, students decode a series of progressively harder pronounceable nonsense words; the TOWRE has a speed component as students are asked to read as many nonwords as possible in 45 seconds. Alternate-form reliability exceeds .91 for Phonemic Decoding Efficiency; test–retest reliability for Word Attack exceeds .73 for this grade-range.

Reading comprehension skills. Measures used are *Gray Oral Reading Test–Version 4* (GORT-4; Wiederholt & Bryant, 2004), *Standardized Reading Inventory-2* (SRI-2; Newcomer, 1999), and *WRMT-R Passage Comprehension* subtest. The GORT-4 and SRI-2 both provide text reading accuracy and comprehension scores. The GORT-4 stories are read aloud once, obtaining a measure of reading rate and comprehension. The SRI-2 stories are read once aloud and once silently, with comprehension measured using lexical, inferential, and factual open-ended questions about the text. Time to read each passage is also recorded to provide an additional indicator of reading rate. The Passage Comprehension task assesses comprehension using a cloze procedure. Test–retest reliability exceeds .85 for the GORT-4, .85 for the SRI-2, and .91 for the WRMT-R.

Spelling. The Peabody Individual Achievement Test-Revised (PIAT-R) *Spelling* subtest assesses the child's ability to recognize standard spellings of spoken words, a measure of orthographic awareness. Test-retest reliabilities were .91 in this age range and with this population (Cirino et al., 2002).

Experimental measures of training and transfer. The intervention participants were tested at pre, mid, post, and follow-up; the control test points were pretest and posttest only. *Sound Combinations* tests the reader's ability to pronounce a set of 30 letter clusters including vowel digraphs (*ee, oa, ai*), diphthongs (*oo, oi, ou*), vowel-controlled consonants (*ge, gi, ce, ci*), *r*- and *l*-controlled vowels, and high frequency bound morphemes (*-ing, -tion*). This measure has been found to be a reliable index of training success (Lovett et al., 1994; Lovett, Lacerenza, & Borden, 2000; Lovett & Steinbach, 1997). Observed internal consistency (Cronbach's alpha) was .83.

The *Challenge Words Test* consists of 55 uninstructed, multisyllabic words that embed the instructed spelling patterns and affixes. This test provides students with the opportunity for application of the decoding strategies taught in both the PHAST and Triple-Focus Programs. It is also a sensitive index of transfer of learning for children and adolescents with RD (Lovett et al., 1994; Lovett, Lacerenza, & Borden, 2000; Lovett & Steinbach, 1997) and consistently produces 70-hr treatment effect sizes ranging from .65 to .85. Observed internal consistency (Cronbach's alpha) was .93.

Word Knowledge Tests (pre- and posttesting only). The *Multiple Definitions* task assesses the student's ability to provide two or more definitions for words with multiple meanings. Students receive credit for each unique definition provided. All items on this test are words presented and discussed in the RAVE-O portion of the Triple-Focus Program, and thus this test serves as a measure of instructed vocabulary content. Test–retest reliability for this task was .66, calculated on a preintervention repeat assessment using a similar sample as reported in Morris et al., 2012.

The WORD Test 2. The *Flexible Word Use subtest* (Bowers, Huisingh, LoGiudice, & Orman, 2004) task, assessing vocabulary knowledge, asks students to produce two meanings for each stimulus word provided. The standardized scoring is 1 or 0 per item—with 1 point only being awarded if the child provides two definitions. Standardized scoring was used, but we also deviated from test administration guidelines and asked participants to provide as many definitions as they could. We then calculated an alternate raw score for this subtest that summed all of the definitions provided, thus yielding similar scoring as used for the Multiple Definitions test but on an uninstructed vocabulary list. The Word Test 2 has average test–retest reliability of .90 and average internal consistency reliability of .81.

Language and cognitive abilities: Predictor variables. The following measures were used as predictor variables.

Phonological processing. We used the *Comprehensive Tests of Phonological Processing* (CTOPP; Wagner, Torgesen, & Rashotte, 1999): (a) *Blending Words* measures the ability to combine orally presented, individual speech sounds into words, and (b) *Elision* measures the ability to repeat a spoken word omitting one of the phonemes. The average internal consistency reliability is .84 for Blending Words, and .89 for Elision.

Naming speed (at multiple time points). The *Rapid Automatized Naming* (RAN; Wolf & Denckla, 2005) tasks assess the ability to rapidly name visual symbols (letters, colors, objects, letters, or combinations). Test–retest reliability exceeds .84. Rapid naming of letters is reported here.

Cognitive ability (pretesting only). The *Wechsler Abbreviated Scale of Intelligence* (WASI; Wechsler, 1999) is an abbreviated measure of verbal and nonverbal cognitive ability, adapted from the *Wechsler Intelligence Scale for Children–III* (Wechsler, 1991), and the *Wechsler Adult Intelligence Scale–III* (Wechsler, 1997). Students were administered all four subtests (Vocabulary, Similarities, Block Design and Matrix Reasoning). Test–retest reliability exceeds .90 for composite scores.

Receptive vocabulary (pretesting). The *Peabody Picture Vocabulary Test—Third Edition* (PPVT-III; Dunn & Dunn, 1997) assesses receptive vocabulary skills; participants select from one of four pictures that which best represents the meaning of a word presented orally. Test–retest reliability exceeds .91 for the PPVT-III.

Visual Sequential Memory (pretesting). The *Visual Sequential Memory* subtest of *The Test of Visual Perceptual Skills-Revised* (Gardner, 1996) tested the child's ability to recall a series of forms just presented from four possible alternatives. Average internal consistency reliability for this age range is .54.

Intervention Conditions

Subjects with similar single word reading levels (WRMT-R Word Identification and Word Attack raw scores) were assigned to an instructional group of four children who received the Triple-Focus Program outlined below. A total of 100–125 intervention sessions were conducted during the school year; children typically were seen in a "pull-out" format for 60 min a day, 5 days a week, in their own schools. At the school's discretion, these intervention sessions were scheduled to occur while their classroom was receiving the day's reading instruction, and this occurred in approximately 75% of cases. Where this schedule was not possible, schools elected to have their children come to the program during art, science, and social science instruction. Classes were not scheduled to occur during math instruction. Because we were in multiple cities, school districts and schools, we chose to allow previous and

current curriculum to vary randomly to better evaluate the generalizability of our specific program results.

The Triple-Focus Program is an experimental reading intervention developed and directly based upon our previous work demonstrating that (a) developmental reading problems are associated with multiple core linguistic and cognitive deficits (phonological awareness, naming speed, and cognitive strategy use) that limit reading acquisition; and that (b) remedial reading interventions that address more then one of these deficits are most effective (Lovett, Lacerenza, Borden, et al., 2000; Morris et al., 2012). The Triple-Focus intervention integrated proven instructional modules from our previous randomized control trial report (Morris et al., 2012).

Samples from the Triple-Focus Scope and Sequence for Lessons 32, 77, and 106 are provided in Table A3 of the Appendix. Looking at lessons sampled from different points in the program illustrates how the components were integrated, the amount of instructional time allocated different components, and how the focus and time allocation shifts over the course of 125 hr of Triple-Focus intervention.

The Triple-Focus Reading Program is an integration of the PHAST Reading Program: Parts One (Decoding) and Two (Comprehension) with the RAVE-O Program (Retrieval, Automaticity, Vocabulary Elaboration, Orthography; Wolf, Miller, & Donnelly, 2000; Wolf et al., 2009). The PHAST Reading Program: Decoding teaches children five specific metacognitive word identification strategies so they may become competent and independent readers (overview in Lovett, Lacerenza, & Borden, 2000). Part Two of the program teaches children comprehension strategies (predicting, summarizing, clarifying, questioning) using a metacognitive approach to improve text reading and comprehension skills (overviews in Lovett, Lacerenza, Steinbach, & De Palma, 2014; Lovett, Lacerenza, De Palma, & Frijters, 2012). The RAVE-O Program is an experimental, fluent comprehension intervention developed by Wolf and her colleagues (Wolf, Donnelly, & Miller, 2000; Wolf et al., 2009) that is based on theoretical neurocognitive models of reading. Specifically, RAVE-O facilitates the development of accuracy and fluency in underlying phonological, orthographic, semantic, syntactic, and morphological skills, and their rapid amalgamation at the sublexical, lexical, and connected text levels. RAVE-O addresses the need to explicitly teach children each of these components, and to teach explicit *interconnections* among these component systems of oral and printed language *at the time* core words are taught (Wolf et al., 2009). Core words are taught that exemplify the polysemous nature of many words, their varied syntactic functions in different contexts, and how morphemes facilitate meaning. Thus, the RAVE-O Program focuses on the linguistic building blocks of reading fluency, as well as three strategies for comprehension.

All groups started at the first lesson, but more advanced groups could progress through the lessons more rapidly. As the program progressed, the number of strategies increased and time was devoted to acquisition of a metacognitive ‘Game Plan’ so that children learned how to select a strategy, monitor its effectiveness, and evaluate the results. The focus moved from building phonological and orthographic skills and knowledge to increasing attention paid multiple components of words and connected text at the semantic, syntactic morpho-syntactic, and discourse levels.

The Triple-Focus Program was designed to teach the children a set of word identification strategies and specific decoding procedures so that they become more competent and independent in their approach to reading unfamiliar words in print—and, at the same time, to develop accuracy and fluency in underlying linguistic retrieval skills so the children could learn to read text fluently and with comprehension. As an extension of two reading interventions that had been used with positive results in our Atlanta, Boston, and Toronto sites, the Triple-Focus program was designed to offer a structured and scaffolded instructional framework of effective decoding and reading strategies. The original five decoding strategies of the PHAST Program were supplemented by and tied to the fluency, orthography, vocabulary, syntax, and morphology activities of the RAVE-O Program. This allowed a richer linguistic framework of component skills and strategies with which to remediate the multifaceted language-based deficits of these struggling readers.

As in our previous implementations, the program began with phonological remediation, acquisition of letter- and letter-cluster sound mappings, phonological analysis and blending skills, and practice using a “sounding out” strategy with precision in how sounds were blended (Engelmann & Bruner, 1988). As strategy-specific preskills and knowledge were acquired, additional word identification strategies were learned and practiced, using a strategy dialogue modeled by the teacher and acquired by the children; these included Rhyming (word identification by analogy- Gaskins et al., 1986), Peeling Off (separating affixes in multisyllabic words), Vowel Alert (learning the multiple pronunciations of vowels and vowel combinations according to their frequency in printed English), and I Spy (useful for compound words—identifying smaller known words). Each lesson contained RAVE-O activities and games, drawing upon words and sublexical patterns from PHAST and the core words of RAVE-O, incorporating words with shared phonemes and orthographic patterns, and semantic richness (multiple meanings) into work on vocabulary and orthographic knowledge, word retrieval, and other linguistic building blocks of reading fluency (Wolf et al., 2000; Wolf & Katzir-Cohen, 2001; Wolf et al., 2009).

The Triple-Focus Program was taught by experienced and certified teachers working for the research teams; some had Masters degrees, and all had special education and/or reading additional qualifications. There were multiple teachers at each site; several had participated in our earlier studies and had experience teaching PHAST and RAVE-O. All teachers were trained to provide multiple interventions within our research (i.e., in other related studies). All were trained a priori to a level of competence during intensive training conducted in Boston. All teachers had a detailed Scope and Sequence, scripted lessons to follow, and timelines with which to adhere for their teaching.

Throughout the study, a senior/lead research teacher at each site (a) continually monitored the progress and pace of each teacher and group through the lessons, (b) initiated cross-site teleconferences between teachers to answer questions and problem-solve challenges, and c) offered reminders and instructional refreshers during regular team meetings. To further support fidelity of implementation in each class and across sites, an email list-serve was established where teachers could ask questions and receive an immediate response. Every teacher prepared a weekly progress report, summarizing all lessons and activities completed by each class. This report was posted weekly on the LISTSERV. In-person mentor visits by the senior/lead teacher occurred 3–4 times per year with feedback being provided. Finally, videotapes of classes

were shared between sites so that trainers/lead teachers could ensure cross-site consistency in program implementation.

The control condition was a curricular control group, including children who met study criteria for RD, and who were not placed in the Triple-Focus intervention; these children constituted a 'business-as-usual' control to be followed and evaluated over time. As a classroom-based control, these children received whatever level and type of intervention the schools or their parents would provide for them. Because schools in these years (2001–2006) frequently waited until 3rd grade to identify children as needing help, it is unlikely that many children in Grades 1 and 2 received extra reading assistance in school. This was the case for all three sites. Schools in all sites provided 90 min of classroom literacy instruction daily. For ethical reasons, these children were offered access to the intervention program the year following their control participation. Control participants were assessed at pre- and posttest only.

The full 125 hr of instruction were implemented as planned for 68% of the sample (n = 117). The remaining 55 intervention children received an average of 104.5 hr of instruction (SD = 14.5; range = 70 to 124 hr). All control participants were assessed on intervention outcomes after equivalent time in the business-as-usual condition. Postintervention assessments for all participants occurred after the final lesson was delivered, with those who did not complete the full 125 hr having their last observation carried forward.

Results

First Analysis

The first analysis is preliminary to the main analyses presented in the following section. The goals of the first analysis were to replicate program findings from our previous work, to generate traditional effect sizes, to maximize power for the intervention versus control contrast, and to provide a conservative test of the efficacy of the intervention. As such, the first analysis included all participants who contributed any valid outcome data, regardless of how much instruction was delivered, carrying forward the last outcome measurement for those who dropped out prior to the planned 125 hr. Moderated regression models were formed, regressing each outcome on preintervention outcome scores, intervention group (i.e., Triple; Control), grade (i.e., a priori focused contrasts of Grades 1/2 vs. 3; Grade 1 vs. 2), the interaction between grade and intervention assignment, and the interaction between preintervention scores and intervention group. This final interaction, representing the homogeneity of regression slopes in ANCOVA, was initially included in each model, but dropped from the final model if nonsignificant. Because each model explicitly included a developmental indicator (i.e., grade), raw scores on each outcome were analyzed, except on the WJ-III outcomes, which utilized the Rasch-scaled W scores.

Because children were nested within their instructional groups, the analysis was conducted within a mixed model framework, with instructional group as a random effect. Cross-level interactions and nested group effects were not of substantive interest and are not reported here, but simply incorporated into each model to account for the group-level dependence in the data and to obtain appropriate standard errors for grade and treatment effects. The resulting intraclass correlations (ICC) are included in Table 2 for use in conducting power analyses for future cluster-randomized trials. In addition to the moderated regressions, intervention effect sizes (Cohen's d) were calculated via the pooled pretest standard deviations of the intervention and control groups, along with the model-adjusted postintervention mean score on each outcome. Table 2 reports these values for outcome models formed for 14 outcomes. Because this analysis involved 14 correlated outcome measures, the potential for an inflated false-discovery rate existed.

Table 2
First Analysis: Results of the Mixed Model Moderated Regressions on Reading and Language Outcomes

Outcome	Model ICC	F^a	Adj. mean difference	SE	CI	d	Grade by treatment	F	p
Experimental measures									
Sound combinations	.43	58.11	9.17	0.96	[7.26, 11.08]	1.81			
Challenge test	.23	70.17	16.65	2.08	[12.52, 20.79]	1.82	2 > 1[b]	4.44	.04
Multiple definitions trained	.34	77.73	0.50	.06	[0.38, 0.62]	1.44	2 > 1	6.05	.02
Standardized/norm-referenced measures									
TOWRE Phonological Decoding	.12	36.30	6.36	1.07	[4.23, 8.49]	1.39			
TOWRE Sight Words	.10	22.69	8.81	1.74	[5.36, 12.26]	0.57	1/2 > 3[c]	3.91	.05
WRAT-3 Reading	.11	15.66	4.40	0.59	[3.23, 5.58]	0.91			
WRMT-R Word Attack	.17	75.32	17.48	1.98	[13.56, 21.41]	1.08	1/2 > 3	3.09	.08
WRMT-R Word Identification	.07	55.34	22.97	2.93	[17.17, 28.77]	0.59	1/2 > 3	3.93	.05
WRMT-R Passage Comprehension	.21	55.69	16.52	2.21	[12.13, 20.91]	0.63			
GORT-R Comprehension	.27	18.23	6.22	1.46	[3.33, 9.11]	0.90			
SRI Comprehension	.27	19.34	7.66	1.74	[4.20, 11.11]	0.64			
GORT-R Rate	.21	23.20	3.86	0.80	[2.27, 5.46]	0.78			
WORD-2	.05	25.06	0.22	0.04	[0.14, 0.31]	0.61			
PIAT-R Spelling	.20	19.23	8.23	1.32	[5.60, 10.84]	0.72	3 > 1/2[d]	7.37	.01

Note. All analyses reported in this table were performed on raw scores, unless the measure was a Woodcock Reading Mastery Test—Revised (WRMT-R) subtest, and these were analyzed using W scores. ICC = intraclass correlation; TOWRE = Test of Word Reading Efficiency; GORT-R = Gray Oral Reading Test–Version 4; SRI = Standardized Reading Inventory; WORD-2 = The WORD Test 2—Elementary; PIAT-R = Peabody Individual Achievement Test—Revised; CI = confidence interval.

[a] Reports the F statistic for the intervention versus control test of posttest adjusted means, all $p < .001$ after adjustment for the false discovery rate. [b] Grade 2 adjusted posttest mean greater than Grade 1. [c] Combined Grade 1/2 adjusted posttest mean greater than Grade 3. [d] Grade 3 adjusted posttest mean greater than combined Grade ½.

Table 3
Second Analysis: Growth Curve Model Fixed Effects for Intervention Group, Grade, and Group by Grade Interactions

Fixed effects	Par.	CHT	WAT	WID	WPC	SRI	TSW	TPD	GRT
Pretest									
Intercept	γ_{00}	4.25**	456.00**	404.05**	439.57**	7.55**	19.31**	4.08**	3.94**
Intervention	γ_{01}	0.33	6.27**	9.64**	6.37**	0.46	2.60	0.72	0.65
Grade 1/2 vs. 3	γ_{02}	−4.08**	−6.49**	−17.46**	−10.66**	−4.57**	−8.01**	−1.70**	−2.89**
Grade 1 vs. 2	γ_{03}	−0.27	−4.84**	−17.38**	−9.70**	−3.21**	−6.31**	−0.91*	−0.83**
Growth to posttest									
Intercept[a]	γ_{10}	21.87**	25.95**	47.33**	29.76**	12.59**	18.64**	6.21**	7.00**
Intervention	γ_{11}	18.15**	15.80**	19.37**	14.80**	8.61**	9.56**	6.21**	4.23**
Grade 1/2 vs. 3	γ_{12}	−0.85	5.81**	11.79**	5.41**	0.76	2.98**	0.23	0.02
Grade 1 vs. 2	γ_{13}	−4.80**	3.99**	15.12**	5.50**	−0.12	3.99**	0.28	−0.07
Interv. × Grade 1/2 vs. 3	γ_{14}	0.46	1.18	5.62*	1.30	0.58	2.44*	0.40	0.03
Interv. × Grade 1 vs. 2	γ_{15}	−6.88*	0.21	4.98	2.74	−0.92	1.92	1.88	0.66
Follow-up trajectory									
Intercept[a]	γ_{20}	2.93*	−0.69	7.19**	4.21**	3.12**	5.81**	2.38**	2.89**
Grade 1/2 vs. 3	γ_{21}	0.86*	−0.05	−0.62	−0.28	0.22	−0.31	−0.54**	−0.07
Grade 1 versus 2	γ_{22}	1.37**	1.32**	0.94	1.56**	0.52	0.68**	0.83**	0.95**

Note. All analyses reported in this table were performed on raw scores, unless the measure was a Woodcock Reading Mastery Test—Revised (WRMT-R) subtest, and these were analyzed using W scores. CHT = Challenge Test; WAT = WRMT Word Attack; WID = WRMT Word Identification; WPC = WRMT Passage Comprehension; SRI = Standardized Reading Inventory Comprehension; TSW = Test of Word Reading Efficiency (TOWRE) Sight Word Efficiency; TPD = TOWRE Phonemic Decoding Efficiency; GRT = Gray Oral Reading Test rate.
[a] All effects FDR corrected at $p < .001$.
* $p < .05$. ** $p < .01$.

As a result, the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) was implemented to correct for multiple significance tests and control the false-discovery rate.

Globally, across all outcomes, statistically significant and substantial program effects were observed on adjusted posttest scores. In every case, participants in the Triple intervention outperformed those in the Control condition, with effect sizes ranging from a moderate .57 to a large 1.82. Effect sizes were largest for experimental outcomes assessing directly instructed content and lowest, but still moderate-to-large, for standardized measures of single word identification. Very strong effect sizes were observed for measures of nonword decoding, and strong effect sizes for reading comprehension outcomes. The average effect size across the 14 outcomes was 0.99. The average effect size on standardized measures was .80 and on experimental measures was 1.69. Intraclass correlations ranged from 0.05 to 0.43, with the largest ICCs observed for experimental measures of instructed content and comprehension outcomes.

The grade by treatment interaction was statistically significant for five outcomes, and marginally significant for one. After accounting for pretest scores, the difference between intervention and control for Grade 1/2 participants at posttest was approximately twice as large as the difference for Grade 3 participants. This pattern was repeated across TOWRE Sight Words (Grade 1/2 intervention-control posttest difference of 11.1 vs. Grade 3 difference of 3.7; illustrated in Figure 1), WRMT-R Word Identification (Grade 1/2 intervention-control difference of 26.6 vs. Grade 3 difference of 14.6), and WRMT-R Word Attack (Grade 1/2 intervention-control difference of 19.7 vs. Grade 3 difference of 12.1; marginally significant). A reverse pattern was observed for three other outcomes, whereby the difference between intervention and control for participants in higher grades was approximately twice as large as the difference for younger participants. The Grade 2 intervention-control difference was greater than the Grade 1 difference on the Challenge test outcome (Grade 2 intervention-control posttest difference of 22.9 vs. Grade 1 difference of 12.2). A similar pattern was observed for The WORD Test 2 outcome (Grade 2 posttest difference of 0.71 vs. Grade 1 difference of 0.36). This pattern was repeated for the PIAT Spelling outcome, with three times the intervention-control posttest difference for Grade 3 participants (15.5) compared with Grade 1/2 participants (5.1).

The interaction between pretest scores and intervention condition was significant for the Sound Combination and PIAT Spelling outcomes. Within the ANCOVA framework, this would indicate a violation of the homogeneity of regression assumption, being differential adjustment of posttest scores by group. Within the moderated regression framework, these two effects can be explicitly modeled and interpreted as substantive effects. In the case of Sound Combinations, a dramatic increase in intervention group score variance from pretest to posttest resulted in a lower pre–post correlation for that group compared ($r = .30$) to Controls ($r = .62$). A similar, but less dramatic, pattern was observed for PIAT Spelling, with a few Intervention participants making large gains by posttest, thereby reducing the correlation between pre- and posttest for the Intervention group.

Second Analysis

The goals of the second analysis were to utilize all available repeated observation data to precisely estimate specific trajectories of change, incorporating trajectories representing the yearly follow-up outcome measurement, and exploring predictors of these two intra-individual parameters. Growth curves were used to estimate intercepts and trajectories, and to model individual differences in intervention response across five repeated observations: pretest, after 35, 70, 105, and 125 hr of instruction, and at each follow-up occasion, which occurred yearly, one to three times after the intervention depending on grade at entry to the study (i.e., up to the end of 4th grade only).

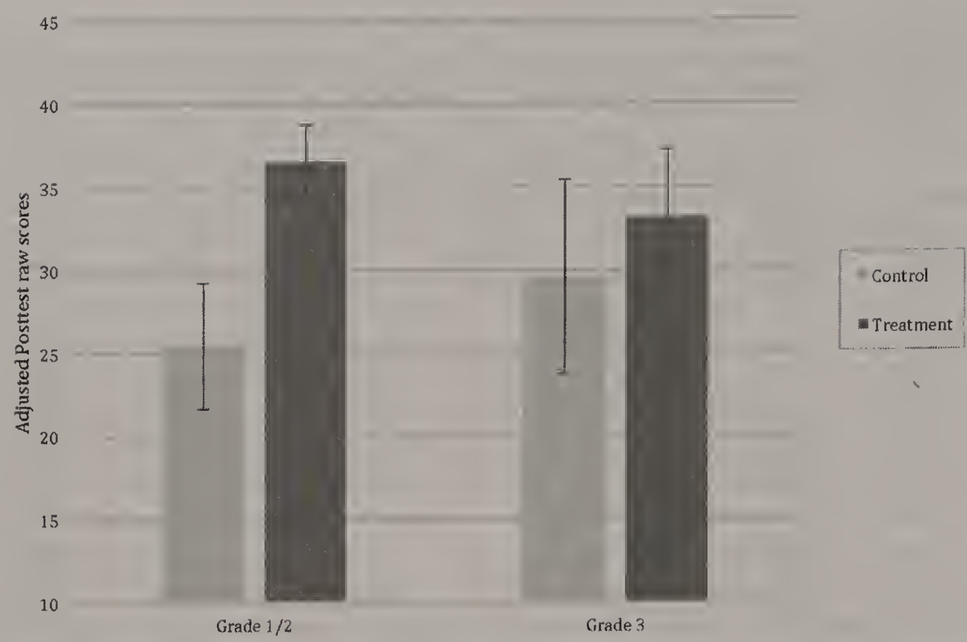


Figure 1. Adjusted posttest means (95% confidence intervals) on Test of Word Reading Efficiency sight words (raw scores).

The growth curve analysis was performed on the eight outcomes for which there were outcome measurements at each of the time-points mentioned above. For precision in estimating fixed effects and to utilize all available data, this second analysis included all cases that had at minimum pretest outcome scores. All available outcome data were utilized, and every dropout was represented in the growth curve analysis. Data density across observation points was as follows: pretest, 219 participants (100%); 35 hr, 172 (78.5%); 70 hr, 172 (78.5%); 100 hr, 158 (72.1%); 125 hr, 205 (93.6%); first follow-up, 52 (23.7%); second follow-up, 52 (23.7%); third follow-up, 28 (12.7%). Lower data density at the 35, 70, and 100 hr testing points and at follow-up represent the fact that control participants were only tested at pretest and posttest.

Growth models were based on a two-piece parameterization of time that modeled linear growth to posttest, with a separate component representing linear growth from posttest through the follow-up period. Not presented here are the competing piecewise and polynomial models that each provided a less adequate fit (via nested −2LL comparisons) across all outcomes. The most notable practical advantage of the chosen parameterization was that it afforded the ability to segregate and estimate effects that might interact with intervention condition from effects that might predict follow-up trajectories. This segregation was important, because follow-up trajectories could not be estimated for participants in the control condition. Several metrics for time were considered, including time as intervention days, chronological time, and models that incorporated both metrics. The most parsimonious and well-fitting model, used in the analyses reported here, was a hybrid model in which intervention days were utilized as the metric for time, with time to follow-up rescaled to this metric.

Model fitting proceeded according to current best practices in multilevel growth modeling (e.g., Hox, Moerbeek, & van de Schoot, 2010; Snijders & Bosker, 2012). Initial model fitting also investigated several models to account for nesting of observations within individuals, within teacher and/or instructional group. Because the control condition was business-as-usual, and thus individual children were not cluster-randomized, clusters of one were formed for analysis purposes. Simulation studies have shown this strategy to be both more efficient and powerful than either forming pseudoclusters, or treating the entire condition as one cluster (Bauer, Sterba, & Hallfors, 2008; Roberts & Roberts, 2005). Null and growth models involving only random effects were first fit and competitively evaluated within measure via BIC/AIC values. Across the eight outcomes, the best fitting random effects model included variance components for intercept and intervention growth rate, both at the participant and participant nested within instructional group levels. The best fitting model per outcome also included the follow-up trajectory piecewise parameter as a variance component, but in almost every case only for participants nested within instructional group. Finally, individual differences were incorporated as predictors of either program-related growth and/or as predictors of follow-up trajectories. The initial fixed-effect predictor model included intervention group, grade at intervention start, and the interaction between group and grade. Fixed effects results from these models are presented in Table 3.

Examination of the fixed effects for Pretest, rows γ_{01} to γ_{03} in Table 3, indicate that participants in Grade 3 began intervention with substantially higher scores on all outcomes when compared with those in Grade 1/2 (row γ_{02}); participants in Grade 2 began with higher scores than those in Grade 1 (row γ_{03}), except for scores on the Challenge Test. When considering Growth to Posttest, row γ_{10} represents the growth rate in the control group, with γ_{11} growth made by Intervention participants over and above this baseline. Given the scaling of the growth models' time parameter, the estimates in these rows are a direct representation of estimated growth over the course of 125 hr of instruction. For example, control participants gained an average of 21.87 Challenge Test words over 125 hr, and Intervention participants gained an *additional* 18.15. Across all outcomes, additional gains by Intervention participants were both substantial and statistically significant.

Rows γ_{12} and γ_{13} represent growth rates across the grade contrasts. On four of eight outcomes, Grade 1/2 participants gained

skills at a faster rate than Grade 3 participants (row γ_{12}). In the case of two outcomes (WRMT-R Word Identification and TOWRE Sight Words), this effect interacted with intervention group (row γ_{14}). In these cases, the growth rate of Grade 1/2 participants in the intervention group far exceeded the rate of growth for Grade 3 participants (see Figure 1). This replicates similar effects seen in the first analysis. On four of eight outcomes, Grade 1 participants gained skills at a faster rate than those in Grade 2 (row γ_{13}). This pattern was reversed for the Challenge Test and interacted with intervention group, such that the intervention effect was much more pronounced for participants who received the intervention in Grade 2 (row γ_{15}).

A pseudo-R^2 (Hox, 2010; Raudenbush & Bryk, 2002) was calculated as an estimate of the proportion of variance in growth rates that could be accounted for by a) assignment to intervention or control condition; and b) the incremental proportion of variance explained by the grade by intervention interaction. Treatment assignment accounted for an average of 32% of the explainable variation in growth rates (range = 21% to 49%); grade by intervention interactions accounted for an average of 52% additional variance in growth rates (range = 29% to 76%).

The rows γ_{20} to γ_{22} in Table 3 characterize follow-up trajectories. Overall, participants continued to gain reading skills from posttest through the follow-up occasions, on all outcomes except WRMT-R Word Attack (row γ_{20}). The parameters in this row represent skill growth per year of follow-up. On six of eight outcomes, continued growth interacted with grade at intervention. In these cases, participants who began the intervention in earlier grades continued gaining skills at a rate that exceeded later intervention starts through the follow-up years (row γ_{22}). Figure 2 illustrates this effect on the WRMT-R Passage Comprehension outcome. Note that by the second follow-up observation, participants starting the intervention in Grade 1 had caught up to those starting the intervention in Grade 2, despite being one year younger at that observation point.

Examination of variance component residuals indicated that additional intra-individual variability remained after accounting for intervention and grade effects. As a result, a secondary analysis was conducted incorporating additional individual difference factors as follows: receptive vocabulary scores (Peabody Picture Vocabulary Test), phonological awareness (Comprehensive Tests of Phonological Processing phonological composite score), rapid naming (Rapid Automatized Naming Letters score), visual sequential memory (Test of Visual Perceptual Skills-Revised), and IQ (Wechsler Abbreviated Scales for Intelligence: 4-subtest IQ score). Each of these factors was introduced to the model initially alone, as predictive of growth to posttest, and as predictive of follow-up trajectories. Interactions of intervention group with these predictors were also included. Models were pruned of higher-order nonsignificant results to reflect a parsimonious model of individual differences. Table 4 reports significant results for these fixed effect individual difference predictors and their interaction with intervention condition.

In Table 4, rows γ_{00} to γ_{04} indicate whether each individual difference predictor was related to pretest scores on each outcome measure. Across multiple outcomes, phonological awareness and rapid automatized naming were related to pretest reading skill. Rows γ_{10} to γ_{14} indicate whether initial levels of the individual difference predictors were related to rate of change during the intervention period. Most of these effects are not interpretable, because they were included to ensure that all nested effects within a significant higher-order interaction were included, as reported in rows γ_{15} to γ_{19}.

Across seven out of eight outcomes, IQ interacted with intervention group growth rates (this relationship was marginally significant for the remaining outcome-GORT Rate). Post hoc examination of these interactions indicated that intervention growth rates were highest among lower-IQ Triple participants, with the greatest discrepancy in growth rates between intervention and control occurring when WASI IQs were low. In fact, the only group not demonstrating growth during the intervention period was the control subgroup with lower WASI IQ scores at entry. Post hoc examination indicated parallel slopes across participants with lower versus higher IQs *if* they participated in the Triple-Focus intervention. The interaction between WASI IQ and response to intervention on the WRMT-R Word Attack outcome is depicted in Figure 3. A reverse pattern was observed on SRI comprehension

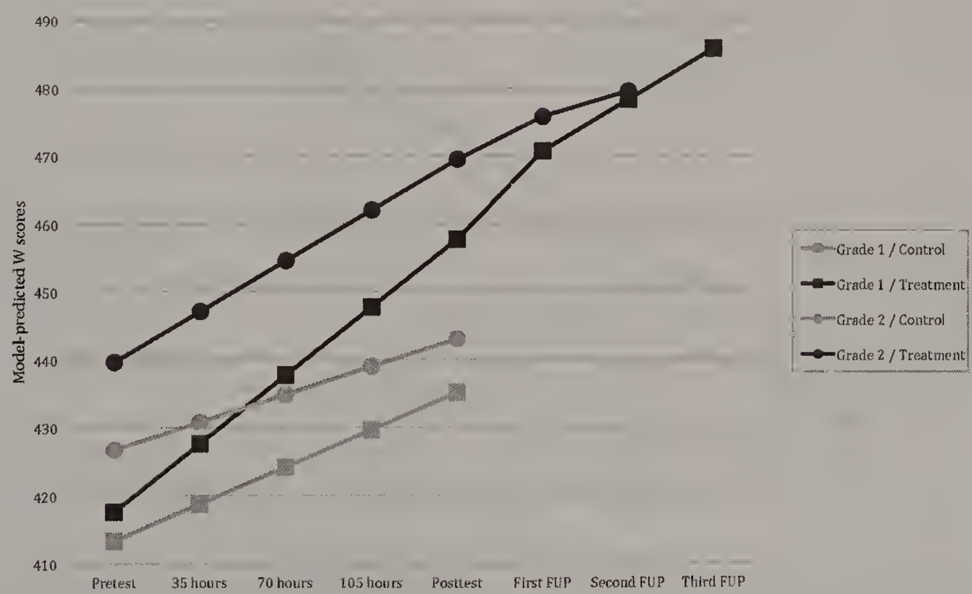


Figure 2. Model-predicted Woodcock Reading Mastery Test—Revised Passage Comprehension (W scores) by grade and intervention status.

Table 4
Second Analysis: Individual Difference Effects and Interactions With Intervention Condition

Fixed effects	Par.	CHT	WAT	WID	WPC	SRI	TSW	TPD	GRT
Pretest									
Phonological awareness	γ_{00}	1.70**	4.04**	4.53**	2.37**	0.63	2.00**	1.53**	0.26
Rapid naming	γ_{01}	2.04**	2.67**	8.44**	6.82**	2.10**	5.71**	0.95**	2.46**
Vocabulary	γ_{02}	−0.54	−1.07	1.49	−0.34	0.56	−1.25*	−0.33	−0.01
Visual sequential memory	γ_{03}	−0.90*	−0.54	0.52	0.83	−0.03	−0.59	−0.56**	−0.01
IQ	γ_{04}	1.16	1.47	1.98	1.53	0.93	1.00	0.48	0.09
Growth to posttest									
Phonological awareness	γ_{10}								
Rapid naming	γ_{11}				−2.62**				
Vocabulary	γ_{12}				2.55**	3.43**			
Visual sequential memory	γ_{13}								
IQ	γ_{14}	0.51	−0.09	1.83	0.01	−0.31	1.04	0.71	−0.04
Int. × Phonological Awareness	γ_{15}								
Int. × Rapid Naming	γ_{16}								
Int. × Vocabulary	γ_{17}					−3.08*			
Int. × Visual Seq. Memory	γ_{18}								
Int. × IQ	γ_{19}	10.79**	16.93**	16.36**	11.11**	7.76*	7.58*	6.02*	1.67†
Follow-up trajectory									
Phonological awareness	γ_{20}						−0.89**	−0.63**	−0.12*
Rapid naming	γ_{21}						0.48*	0.66**	0.36*
Vocabulary	γ_{22}			2.14**	0.60†	1.95**	1.99**	1.34**	0.33**
Visual sequential memory	γ_{23}	1.58**	0.85**					0.36*	1.12**
IQ	γ_{24}	−1.02*	−1.46**	−1.92*	−1.24**				

Note. All analyses reported in this table were performed on raw scores, unless the measure was from a Woodcock Reading Mastery Test—Revised (WRMT-R) subtest, and these were analyzed using W scores. CHT = Challenge Test; WAT = WRMT Word Attack; WID = WRMT Word Identification; WPC = WRMT Passage Comprehension; SRI = Standardized Reading Inventory Comprehension; TSW = Test of Word Reading Efficiency (TOWRE) Sight Word Efficiency; TPD = TOWRE Phonemic Decoding Efficiency; GRT = Gray Oral Reading Test rate; Int. = Intervention.
† $p < .10$. * $p < .05$. ** $p < .01$.

outcomes with receptive vocabulary (PPVT) as the predictor (row γ_{12}). Higher vocabulary scores were associated with a greater difference in growth rates between intervention and control participants. The greatest growth rates in comprehension were observed for Triple-Focus participants who began intervention with relatively stronger vocabulary skills.

Rows γ_{20} to γ_{24} indicate the relationship between growth trajectories during the follow-up period and the individual difference predictors. Across four of eight outcomes (row γ_{23}), higher visual sequential memory skill was associated with greater gains in the follow-up period. Post hoc visual inspection of this effect showed that participants with the highest visual sequential memory skills continued to gain reading skills, whereas those with the lowest did not continue to gain, but rather leveled off one year after intervention. This pattern is depicted in Figure 4 for the measure of multisyllabic challenge word reading. The same pattern was evi-

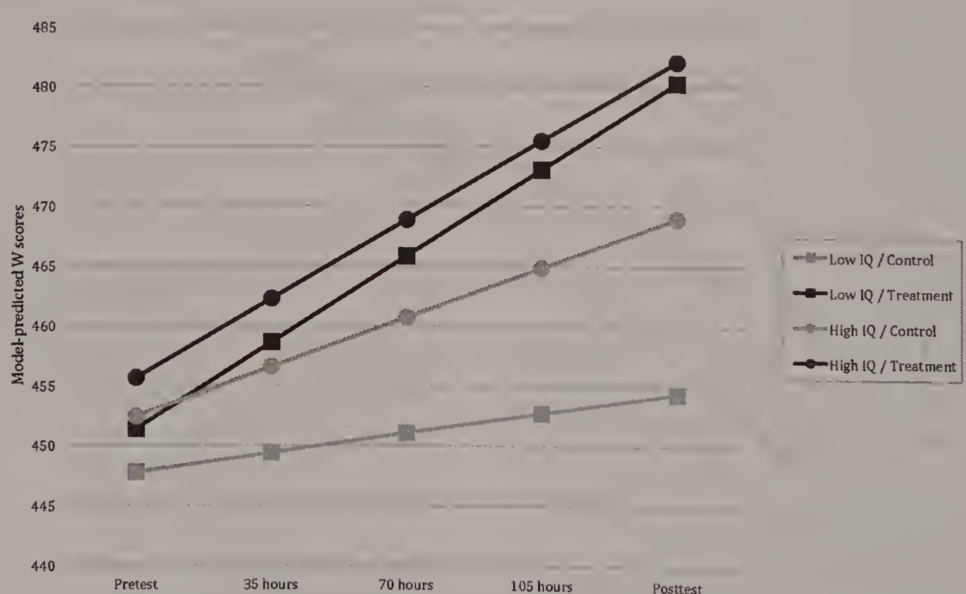


Figure 3. Model-predicted Woodcock Reading Mastery Test—Revised Word Attack (W scores) by low versus high Wechsler Abbreviated Scale of Intelligence IQ scores and intervention status.

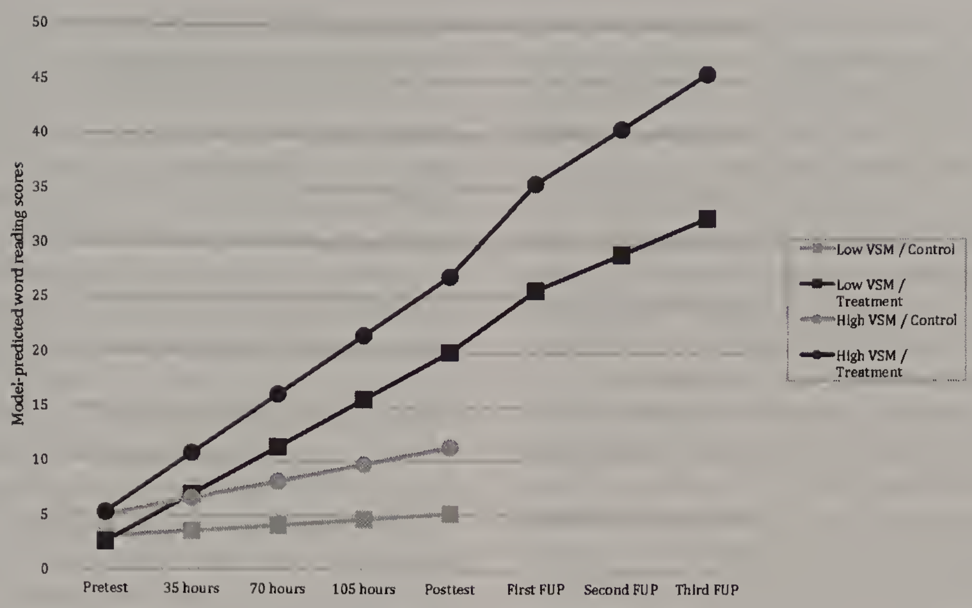


Figure 4. Model-predicted multisyllabic challenge word reading scores by low versus high visual sequential memory scores and intervention status.

dent on five of eight outcomes for the receptive vocabulary predictor. Relatively higher vocabulary skill was associated with greater continued gains during the follow-up period (row γ_{22}).

Normalization Rates

A final examination was made of the proportion of participants in each condition and each grade whose posttest scores fell within the average range following the intervention period. The proportions 'normalized' on four standardized outcome measures are displayed in Table 5.

Chi-square tests of independence were calculated to establish whether the proportion of children achieving scores in the average range at posttest differed between the Triple and control groups. On the WRMT-R subtests (Word Attack, Word Identification, Passage Comprehension), significantly greater normalization was achieved by the Triple group in every grade, the only exception being a greater but nonsignificant advantage of the Triple over the control children in Grade 3 on Word Identification. On the SRI-2, lower rates of normalization were observed overall, however, Triple intervention children were normalized at significantly greater rates for SRI Accuracy and Reading Quotient scores in Grades 1 and 2, but the difference fell short of significance for Grade 3 children.

Discussion

The preliminary analysis confirmed that the research intervention, the Triple-Focus Reading Program, was associated with reliable gains in reading achievement that were evident on multiple dimensions of reading skill. Across 14 reading outcomes, ranging from experimental measures of skills targeted for instruction (e.g., Sound Combinations, multisyllabic Challenge Word reading, Multiple Definitions vocabulary knowledge) to standardized measures of word identification, word attack, word reading efficiency, and reading comprehension, children who received the Triple-Focus intervention substantially out-performed those in the control condition. Effect sizes (Cohen's *d*) ranged from .57 to 1.82, with an average effect size of 0.99 and a median effect size of .84. These effect sizes, achieved after only 125 hr of instruction (approximately 7 months chronological time), are comparable to those reported by Connor et al. (2013) comparing three years of ISI

Table 5
Percentage of Participants Normalized (>90 SS by Final Posttest) Across Four Reading Outcomes

	Triple intervention			Control		
Outcome	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
WRMT-R Word Identification	76.3*	52.6*	21.3	38.9	7.7	12.5
WRMT-R Word Attack	77.6*	50.0*	38.3*	27.8	0.0	6.3
WRMT-R Passage Comprehension	67.1*	36.8*	34.0*	29.4	0.0	6.3
SRI Passage Comprehension	40.3	21.6	28.3	17.6	0.0	12.5
SRI Accuracy	61.1*	62.2*	57.8	17.6	7.7	37.5
SRI Reading Quotient	40.3*	24.3*	15.6	11.8	0.0	6.3

Note. WRMT-R = Woodcock Reading Mastery Test—Revised; SRI = Standardized Reading Inventory.
* Chi-square test indicated that the proportion normalized within grade differed across intervention and control conditions, $p < .05$.

intervention to three years of control placement. The present effect sizes surpass most of those reported in the meta-analysis conducted by Wanzek and Vaughn (2007), however, in which effects were generally greater for children receiving intervention in Kindergarten and Grade 1 (average e.s. ranging from .31 to .84) than in Grades 2 or 3 (.23–.27).

The efficacy of the Triple-Focus Reading intervention was anticipated because its component programs (PHAST and RAVE-O) had been rigorously evaluated against two control groups in a previous multisite study with Grades 2 and 3 children with reading disability (Morris et al., 2012). Both multiple component programs shared an emphasis on phonology, orthography, and morphology, and both included specific motivational and metacognitive components in their design. The two programs offered the same base of phonological reading intervention (PHAB/DI), but differed in some other areas of cognitive-linguistic focus. The RAVE-O Program provided instruction on several linguistic aspects of word knowledge (e.g., semantic depth and flexibility, lexical retrieval, syntactic and morpho-syntactic structure) and offered many game-like practice opportunities to build engagement with language learning. The PHAST program provided a metacognitive approach to decoding, with attention paid subsyllabic orthographic patterns, variable vowel pronunciations, affixes, and the direct teaching of five word identification strategies, along with a plan for their implementation, monitoring, and evaluation. These two research-based intervention programs were associated with superior outcomes relative to controls on multiple standardized reading achievement tests at posttest, and participants continued to demonstrate a significant advantage a full year after intervention ended (Morris et al., 2012). The superiority of the PHAST and PHAB/DI + RAVE-O programs was replicated across multiple measures of reading and spelling achievement at 1-year follow-up. In the present design, these multidimensional programs were integrated and extended to form the Triple-Focus program, and given our previous evidence, there was ample reason to believe that the new intervention would have efficacy for struggling readers in the early grades.

In this previous work, interventions offered only 70 hr of small-group instruction. Sampling was conducted according to a 2 × 2 × 2 factorial design such that every treatment group included equal numbers of Caucasian and Black children, children from average or below-average family socioeconomic circumstances, and children of average or below-average IQ (IQs 70–89). In this study, program benefits generalized to a much broader sample of disabled readers than typically evaluated. These multidimensional, systematic and intense, linguistically motivated reading interventions were associated with positive outcomes for young children with RD, of high and low IQ, and from a range of ethnic backgrounds and environmental circumstances (Morris et al., 2012).

In the present research, 125 hr of small-group instruction was offered over the course of 1st, 2nd, or 3rd grade, allowing an integration of the PHAST and RAVE-O components and further development of reading comprehension instruction. Of primary interest was the question of whether grade at intervention would influence intervention outcomes and rate of growth. Although robust intervention effects were observed on all 14 outcomes, the interaction between intervention condition and grade-at-intervention was significant for just less than half of these outcomes: For five of the 14 outcomes, outcomes differed according to grade and a 6th outcome was marginal. In five of six cases, these effects concerned acquisition of basic foundational reading skills.

There was powerful evidence of an early intervention advantage on most basic word reading skills assessed: For word attack, word identification (WRMT-R), and sight word reading efficiency (TOWRE), intervention in Grades 1 and 2 was associated with greater gains than in Grade 3. The only standardized word reading measure that did not demonstrate this advantage was WRAT-3 Reading. Phonological decoding training appeared to benefit all grades equally on measures of letter-sound combination knowledge and nonword reading efficiency (TOWRE decoding), although on one central measure of decoding skills (WRMT Word Attack), 1st and 2nd Grade Triple children were at a substantial advantage relative to 3rd Grade Triple children.

After controlling for pretest, the average posttest difference between intervention and control Grade 1/2 participants was 20.6 W-scores, relative to 12.1 for Grade 3 participants. On Word Identification (28.7 vs. 14.5 W-scores) and TOWRE sight words outcomes (11.9 vs. 3.9 words), the two younger grades demonstrated posttest advantages relative to controls two to three times as great as those in Grade 3. These grade-by-intervention interaction effects were substantial: after accounting for the effects of assignment to condition, grade-by-intervention effects accounted for an average of 54% of the explainable variation in growth rates.

These data provide evidence in support of the efficacy of early intervention within the 1st or 2nd grade of elementary school. This result is of practical significance given the still prevailing stance of some school districts to delay detailed assessment until a child reaches 3rd grade with persisting academic problems. These results are consistent with those of Connor and colleagues (Connor et al., 2013) who found a 1st grade advantage for students receiving only one year of her ISI intervention relative to those whose single year of ISI occurred in 2nd or 3rd grade. Connor et al. noted, however, that their 1st grade advantage was inconsistent and not replicated for students receiving two years of ISI intervention. In this case, students receiving ISI in 1st and 3rd grades outperformed those with ISI in 1st and 2nd or 2nd and 3rd grades.

In the present data, the early intervention effect was not replicated on three other outcomes for which a significant Intervention × Grade interaction was revealed. As the word literacy outcome became more complex, different Grade × Intervention patterns emerged. On two measures relevant to specific metacognitive and metalinguistic instruction in the Triple-Focus Program, multisyllabic word identification (Challenge Words) and the ability to provide multiple definitions of multiple-meaning vocabulary (Multiple Definitions), 2nd Grade Triple children demonstrated a greater intervention-control posttest advantage than 1st Grade Triple children. Finally on an orthographic awareness or spelling recognition measure (PIAT Spelling), 3rd Grade Triple participants achieved a greater posttest advantage relative to controls than 1st and 2nd Grade Triple participants. Each of these outcome measures requires awareness of and a capacity to manipulate linguistic components of written language beyond the phonological domain. On Challenge Words, morphological awareness and an ability to work with bound morphemes are tapped, on Multiple Definitions, morpho-syntactic and semantic awareness and flexibility, and on PIAT Spelling, orthographic awareness. These findings are among the first to attempt to examine developmental

effects in disabled readers' response to intervention according to the complexity of the component reading skills being assessed.

Perhaps the most complex aspects of reading development involve the comprehension of connected text. In this regard, it is of interest that no Intervention Condition × Grade interactions were found on any of the three standardized reading comprehension tests included in the pre- and posttest battery. Substantial intervention effects were revealed on all three comprehension outcomes, with large effect sizes reported (GORT Comprehension $d = .90$, SRI Comprehension $d = .64$, WRMT-R Passage Comprehension $d = .63$). Similarly, a measure of text reading rate (GORT Rate $d = .78$) demonstrated a reliable posttest advantage for the Triple intervention participants, but no interaction with grade.

It is difficult to know whether this pattern truly reflects no grade differences in how the Triple intervention affected reading comprehension performance. The failure to observe developmental response differences on these text reading measures may reflect instead the current state of measurement for more complex dimensions of reading skill like text comprehension and reading fluency. It is acknowledged that traditional measures assess somewhat crudely the *products* of reading comprehension—what is understood after a text is read. Different limitations of these standardized reading comprehension tests have been extensively discussed, including inadequate content validity, concurrent validity, task sensitivity, and an imbalance in the type of comprehension questions included (Cain & Oakhill, 2006; Cutting & Scarborough, 2006; Keenan, Betjemann, & Olson, 2008; Kendeou, Papdopoulos, & Spanoudis, 2012; Morsy, Kieffer, & Snow, 2010).

These concerns are particularly acute when attempting to assess comprehension of text by beginning readers whose skills are undergoing rapid developmental change. Kendeou and colleagues reported a longitudinal study comparing different comprehension measures widely used in the early grades (Kendeou et al., 2012). These investigators demonstrated that these tests vary in the processing demands they make on young readers' component reading-related skills (e.g., vocabulary, orthographic processing, rapid naming, phonological processing, working memory, fluency), skills that are developing rapidly during the early grades. Of relevance to the present work, one of the tests used here as an outcome measure, Passage Comprehension, was found to exert particular processing demands on orthographic processing and working memory, and less on phonological decoding. It should be noted, however, that the Kendeou study sample included typically developing Greek children, and that phonological decoding is typically mastered early in reading development in languages like Greek with highly consistent letter-sound mapping.

The growth curve analyses undertaken with the present data allowed two types of determination of individual differences effects: a) an examination of predictors of growth during the intervention period, with Triple-specific determinants of growth revealed through interactions between the predictor and intervention condition (Triple vs. control); and b) predictors of growth following intervention for children who had received the Triple. Control children did not contribute to the follow-up data because after posttest they received the Triple reading intervention.

Compatible with evidence establishing phonological awareness and rapid naming speed as predictors of reading achievement in young readers (Kirby, Parrila, & Pfeiffer, 2003; Manis, Doi, & Bhadha, 2000; Parrila, Kirby, & McQuarrie, 2004), these two factors were related to pretest reading skill for the present sample. These individual difference factors were not, however, consistently related to rate of growth during intervention or over the follow-up years.

In contrast, and across seven of the eight outcomes analyzed, IQ interacted with intervention condition on rates of growth during the intervention period. Specifically, better rates of growth in the Triple intervention, relative to controls, were seen among those participants with lower WASI IQ estimates. Another way of expressing this interaction is that the difference in growth rate between Triple and Control children was greatest for children with lower WASI scores at entry. This is put into context by the fact that Control participants with lower WASI IQs did not demonstrate growth across the intervention period in marked contrast to higher WASI Control children (see WRMT-R Word Attack growth, Figure 3). Post hoc examination revealed parallel slopes (rates of growth) across Triple participants with higher- and lower WASI scores, indicating that equal growth was attained in the intervention for lower and higher-IQ children. These data suggest that the provision of systematic, linguistically informed, and intense reading intervention is particularly critical for struggling readers with lower overall cognitive and language functioning, and that children of varying cognitive profiles at entry were able to profit equally from the Triple instruction.

Two other individual differences factors emerged to be of interest. The first was an estimate of vocabulary knowledge (PPVT), and this factor was associated with a different pattern than that seen for WASI IQ. In this case, the difference in growth rates between Triple and control children was greater for those children demonstrating relatively stronger vocabulary skills at entry. Greater growth in intervention on SRI Comprehension was observed for Triple children with relatively better vocabulary skills. Similarly greater continued growth on four outcomes including single-word reading, comprehension, and fluency during the follow-up period was revealed for high-vocabulary Triple children. It is not surprising that vocabulary knowledge is related to growth in and development of comprehension skills; there is evidence of the substantial correlations between estimates of vocabulary knowledge and reading ability (Baumann, Kame'enui, & Ash, 2003; Kamil, 2004; Nagy, 2007).

An unexpected predictor of growth during the follow-up period emerged for four of eight outcomes. Visual Sequential Memory performance predicted rate of growth after intervention ended on four reading outcome measures over the follow-up period. Triple children with higher visual sequential memory scores at pretest made greater continued growth on word attack and nonword reading efficiency standardized measures, on multisyllabic challenge word reading, and on GORT Reading Rate over the follow-up year. Verbal working memory is more typically related to differences in reading growth among children and youth, but recent evidence by Pham and Hasson (2014) suggests that visual spatial working memory also significantly predicts reading achievement in children. Swanson (2000, 2010) has speculated that any advantage that visual spatial working memory may give to children with reading disabilities may vary according to the processing demands that reading places on different components of the working memory system.

Grade at intervention continued to exert an influence in predicting differences in rate of growth during the follow-up period. Although many of the Intervention × Grade interactions revealed in the first analysis of posttest change revealed an advantage for

Triple participants in Grades 1 and 2, a more specific advantage for 1st graders is found in follow-up growth rates. Children who received the Triple intervention in 1st grade continued to grow during the following three years at faster rates than children who received the intervention in 2nd grade. The Grade × Follow-Up effect was consistently found across outcome measures, with six of the eight demonstrating this robust effect.

The latter observation of superior growth *after* treatment for our 1st grade sample is reinforced by examination of the normalization rates achieved on different dimensions of reading development by children receiving intervention in Grades 1, 2, or 3. On standardized tests of word attack, word identification, and passage comprehension, significantly greater normalization was attained by Triple participants in every grade, with the exception of a greater but nonsignificant advantage of Triple over control children in Grade 3 on word identification. On the word attack measure, 78% of Triple 1st graders scored within the average range at posttest, 50% of Triple 2nd graders, and 38% of Triple 3rd graders. In contrast, control participants were normalized at the following rates in these grades: 28% 1st grade, 0% 2nd grade and 6% 3rd grade. Although relatively fewer participants were normalized on the SRI Reading Quotient, 40% of Triple 1st graders and 24% of Triple 2nd graders scored within the average range at posttest compared with 12% and 0% of their control peers.

These data provide further support for the clear benefits of early intervention, particularly in 1st grade. These findings are compatible with those recently reported by Al Otaiba and colleagues (Al Otaiba et al., 2014). These investigators compared two response-to-intervention (RTI) models implemented in 34 first grade classrooms using a randomized controlled design. A typical RTI procedure, that deferred further intervention until Tier 1 response was measured, was compared with a dynamic RTI model that provided Tier 2 or Tier 3 intervention immediately based on children's screening results. The interventions differed only with respect to when intervention began. Children in the dynamic RTI condition had significantly higher reading achievement at the end of 1st grade than children in the typical RTI condition. As with the present results, these findings suggest that delaying intervention for struggling early readers is not associated with any advantage for the children; to the contrary, the best outcomes are seen with intervention that begins in Kindergarten or Grade 1. As Al Otaiba and colleagues indicate, any effect of false negatives seems negligible. And as our follow-up data demonstrate, superior growth for 1st graders on foundational reading skills continues over three years after the intervention ends.

Limitations

Clear limitations characterize the present study and qualify the findings. The most obvious concerns the inability to randomly assign participants to intervention or control conditions. Quasi-experimental research designs lack the credibility of RCTs with respect to assessing causality. An inability to randomly assign participants to treatment and control conditions is not uncommon in clinical and applied research settings however (Gliner & Morgan, 2000; Harris et al., 2006). The use of both repeated measurement and a comparison group makes it easier to avoid certain threats to validity within a quasi-experimental design. In that regard, the present study provided compelling evidence of the comparability of intervention and control groups on all selection criteria, and on all pretest and demographic measures.

Another limitation concerns the unequal sample sizes for the intervention and control groups, and resulting imbalance across intervention and control conditions within each grade. In our study, the lower control group numbers were associated with difficulty enrolling control children with reading disabilities who would be required to wait a full year before receiving the intervention.

Related and equally important issues qualify interpretation of the follow-up data. First, no follow-up data are available for control participants. The failure to follow control children untreated over a follow-up period was due to the ethical need to offer the Triple intervention to control children following their posttest assessment. Although Triple placement could not be arranged for all control children for logistical reasons (school location, transportation, etc.), no follow-up assessment was conducted for them because of the intent to offer intervention. This necessitated two separate analyses to consider predictors of outcomes immediately following intervention and then in the years after intervention ended. Second, decreased numbers were available for follow-up analyses due to attrition of the intervention sample over the follow-up years. Of the intervention participants, follow-up data were collected on 30.2% at Follow-up Year 1, 30.2% at Follow-up Year 2, and 16.3% at Follow-up Year 3. Follow-up was only conducted until the end of 4th grade, and so follow-up opportunities decreased with intervention at later grades. These limitations qualify the conclusions that may be offered on the basis of these follow-up data.

A third limitation also concerns the control comparison available in this curricular control design. Because early intervention and RTI initiatives were less prevalent during the time period of the present study, it is possible that some control participants received less reading instruction overall that those in the intervention group. Some intervention children received whole class reading instruction in addition to the small Group Triple intervention program. In these cases, because small group intervention is considered a good vehicle for intensifying reading instruction for struggling readers, it is difficult to determine whether these intervention participants' postintervention superiority can be attributed to the research-based Triple intervention itself, to the additional amount of reading instruction offered, or to the increased individual attention that small group programs can afford. This concern is attenuated somewhat by two previous RCTs demonstrating efficacy of the components of the Triple-Focus intervention relative to other intervention conditions offering additional reading instruction in groups of equal size (Lovett, Lacerenza, & Borden, 2000; Morris et al., 2012).

In addition, it should be noted that the majority of intervention children (approximately 75%) attended their Triple-Focus class *during* the time of whole class literacy instruction. These intervention sessions were scheduled at the school's discretion and many school boards preferred that classes occur while regular classroom reading instruction was occurring. Where this scheduling was not possible, schools generally elected to have children come to the program during art, science, and social science instruction.

An added limitation relevant to the control participants was the lack of specific information on what type of literacy instruction they received in their schools. The majority of control children (81%) were from Toronto and surrounding area schools, and an eclectic approach to literacy instruction was used, largely at the teacher's discretion. The Ministry of Education in the Province of Ontario provided general instructional guidelines but did not endorse any particular reading program or instructional approach during those years. Children in the elementary grades received 90 min of literacy instruction daily, cov-

ering reading and writing activities, and including a range of approaches. The same 90-min block of literacy instruction also characterized our schools in Atlanta and Boston.

Other limitations relate to the context and time within which these data were collected. The study was undertaken in two American cities (Boston, Atlanta) and one Canadian city (Toronto). The study was conducted during a time when No Child Left Behind (NCLB) legislation in the United States may have affected the instructional practices of teachers in early reading and math instruction, and this may have disproportionately affected control participants from U.S. sites. As noted above, however, 81% of control children were from Toronto schools. Although it could be speculated that Canadian schools were not as influenced by the instructional emphases encouraged by NCLB and therefore at a disadvantage, the superiority of Canadian students' reading achievement results over those of their American peers in international comparisons (PISA) might assuage such concern. Students in Canada outrank those in the United States in reading, math, and science on PISA testing. Canada is ranked 7th in the world, whereas the United States is ranked 24th on PISA reading scores (Organisation for Economic Cooperation and Development, 2013). Although conclusions from this research are necessarily qualified by all these contextual and design factors, it is unlikely that the preponderance of Canadian controls biased intervention findings in a positive direction.

Finally, because this study was conducted in three sites with quite different school calendars, there was a difference between sites in the ability to complete 125 hr of intervention within the school year. The full 125 hr of instruction were implemented as planned for 68% of the intervention sample ($n = 117$). The remaining 55 intervention children received an average of 104.5 hr of instruction ($SD = 14.5$; range = 70 to 124 hr). Data density varied considerably across time-points for testing, therefore, and as expected fewer participants were available for follow-up assessment. As in other long-term studies, attrition during the follow-up years occurred.

Conclusion

In conclusion, the present study contributes more evidence on the relative importance of the timing of early intervention for reading problems in the primary grades. Although the Triple-Focus intervention was associated with benefits for struggling readers across 1st, 2nd, and 3rd grades, on all reading and reading-related outcomes, there was a marked advantage on some outcomes for early intervention. Children who received intervention earlier, in 1st and 2nd grade, made gains relative to control children almost twice that of children receiving intervention in 3rd grade on foundational word reading skills such as word attack, word identification, and sight word efficiency. On follow-up testing, the advantage of 1st grade intervention was even clearer: First graders in the Triple condition continued to grow at faster rates over the follow-up years than 2nd graders on six of eight reading outcomes (word attack, passage comprehension, sight word and phonemic reading efficiency, multisyllabic challenge word reading, and GORT reading rate.). Normalization rates indicated that a majority of first graders in the Triple intervention improved and achieved age-appropriate performance scores at posttest on the WMRT reading achievement subtests. These findings suggest that the cost of investing in first grade intervention, using an instructional vehicle with demonstrated efficacy, is offset by the substantial immediate gains, benefits still evident years after the intervention ends. The substantial effect sizes attained with provision of 100–125 hr of intervention provide compelling evidence for the early intervention position.

Finally, the present study is one of the first to examine grade effects in intervention response according to different types of reading outcomes. Evidence was provided of developmental differences in intervention response according to the complexity of the component reading skills being evaluated. On two measures relevant to metacognitive and metalinguistic aspects of the Triple instruction (Challenge Words, Multiple Definitions), 2nd Grade Triple children demonstrated a greater posttest advantage relative to controls than 1st Grade Triple children. On an orthographic awareness measure (PIAT spelling), 3rd Grade Triple children achieved a greater posttest advantage over controls than the 1st or 2nd grade participants. On these outcome measures that require an ability to manipulate linguistic components of written language beyond the phonological dimension, 2nd and 3rd graders enjoyed some intervention advantage. On tests of reading comprehension, however, despite robust intervention effects and large effect sizes, no intervention-by-grade interactions were revealed. This may be attributable to difficulties in reading comprehension measurement for this age and level of reading skill.

References

Al Otaiba, S. (2000). *Children who do not respond to early literacy instruction: A longitudinal study*. Unpublished doctoral dissertation, Vanderbilt University. Nashville, TN.

Al Otaiba, S., Connor, C. M., Folsom, J. S., Wanzek, J., Greulich, L., Schatschneider, C., & Wagner, R. K. (2014). To wait in Tier 1 or intervene immediately: A randomized experiment examining first grade response to intervention (RTI) in reading. *Exceptional Children, 81,* 11–27. http://dx.doi.org/10.1177/0014402914532234

Al Otaiba, S., & Fuchs, D. (2002). Characteristics of children who are unresponsive to early literacy intervention: A review of the literature. *Remedial and Special Education, 23,* 300–316. http://dx.doi.org/10.1177/07419325020230050501

Al Otaiba, S., & Fuchs, D. (2006). Who are the young children for whom best practices in reading are ineffective? An experimental and longitudinal study. *Journal of Learning Disabilities, 39,* 414–431. http://dx.doi.org/10.1177/00222194060390050401

Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research, 43,* 210–236. http://dx.doi.org/10.1080/00273170802034810

Baumann, J. F., Kame'enui, E. J., & Ash, G. E. (2003). Research on vocabulary instruction: Voltaire redux. In J. Flood, J. M. Jensen, D. Lapp & J. R. Squire (Eds.), *Handbook of research in teaching the English language arts* (2nd ed., pp. 752–785). New York, NY: MacMillan.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B. Methodological, 57,* 289–300.

Berninger, V. W., Abbott, R. D., Brooksher, R., Lemos, Z., Ogier, S., Zook, D., & Mostafapour, E. (2000). A connectionist approach to making the predictability of English orthography explicit to at-risk beginning readers: Evidence for alternative, effective strategies. *Developmental Neuropsychology, 17,* 241–271. http://dx.doi.org/10.1207/S15326942DN1702_06

Berninger, V. W., Abbott, R. D., Vermeulen, K., Ogier, S., Brooksher, R., Zook, D., & Lemos, Z. (2002). Comparison of faster and slower responders to early intervention in reading: Differentiating features of their language profiles. *Learning Disability Quarterly, 25,* 59–76. http://dx.doi.org/10.2307/1511191

Berninger, V. W., Nagy, W. E., Carlisle, J. F., Thomson, J. B., Hoffer, D., Abbot, S., . . . Aylward, E. H. (2003). Effective treatment for children

with dyslexia in Grades 4–6: Behavioral and brain evidence. In B. R. Foorman (Ed.), *Preventing and remediating reading difficulties: Bringing science to scale* (pp. 381–417). Timonium, MD: York.

Blishen, B. R., Carroll, W. K., & Moore, C. (1987). The 1981 socioeconomic index for occupations in Canada. *The Canadian Review of Sociology and Anthropology/La Revue Canadienne de Sociologie et d'Anthropologie, 24,* 465–488. http://dx.doi.org/10.1111/j.1755-618X.1987.tb00639.x

Bowers, L., Huisingh, R., Johnson, P. F., LoGiudice, C., & Orman, J. (2004). *The Word Test 2: Elementary*. East Moline, IL: LinguiSystems.

Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *The British Journal of Educational Psychology, 76,* 683–696. http://dx.doi.org/10.1348/000709905X67610

Cirino, P. T., Chin, C. E., Sevcik, R. A., Wolf, M., Lovett, M., & Morris, R. D. (2002). Measuring socioeconomic status: Reliability and preliminary validity for different approaches. *Assessment, 9,* 145–155. http://dx.doi.org/10.1177/10791102009002005

Compton, D. L., Miller, A. C., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken reading theory in the name of "quick fix" interventions for children with reading disability? *Scientific Studies of Reading, 18,* 55–73. http://dx.doi.org/10.1080/10888438.2013.836200

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science, 24,* 1408–1419. http://dx.doi.org/10.1177/0956797612472204

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., . . . Schatschneider, C. (2011). Testing the impact of Child Characteristics × Instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly, 46,* 189–221.

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007, January 26). The early years. Algorithm-guided individualized reading instruction. *Science, 315,* 464–465. http://dx.doi.org/10.1126/science.1134513

Connor, C. M., Morrison, F. J., Schatschneider, C., Toste, J., Lundblom, E., Crowe, E. C., & Fishman, B. (2011). Effective classroom instruction: Implications of child characteristics by reading instruction interactions on first graders' word reading achievement. *Journal of Research on Educational Effectiveness, 4,* 173–207. http://dx.doi.org/10.1080/19345747.2010.510179

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10,* 277–299. http://dx.doi.org/10.1207/s1532799xssr1003_5

Denton, C. A., Fletcher, J. M., Anthony, J. L., & Francis, D. J. (2006). An evaluation of intensive intervention for students with persistent reading difficulties. *Journal of Learning Disabilities, 39,* 447–466. http://dx.doi.org/10.1177/00222194060390050601

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research, 79,* 262–300. http://dx.doi.org/10.3102/0034654308325998

Ehrhardt, J., Huntington, N., Molino, J., & Barbaresi, W. (2013). Special education and later academic achievement. *Journal of Developmental and Behavioral Pediatrics, 34,* 111–119. http://dx.doi.org/10.1097/DBP.0b013e31827df53f

Engelmann, S., & Bruner, E. C. (1988). *Reading Mastery I/II Fast Cycle: Teacher's Guide*. Chicago, IL: Science Research Associates, Inc.

Entwisle, D. R., & Astone, N. M. (1994). Some practical guidelines for measuring youth's race/ethnicity and socioeconomic status. *Child Development, 65,* 1521–1540. http://dx.doi.org/10.2307/1131278

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). *Learning Disabilities: From Identification to Intervention*. New York, NY: Guilford Press.

Fletcher, J. M., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives, 3,* 30–37. http://dx.doi.org/10.1111/j.1750-8606.2008.00072.x

Foorman, B. R., & Al Otaiba, S. (2009). Reading remediation: State of the art. In K. R. Pugh & P. McCardle (Eds.), *How children learn to read: Current issues and new directions in the integration of cognition, neurobiology and genetics of reading and dyslexia research and practice* (pp. 257–274). New York, NY: Psychology Press.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90,* 37–55. http://dx.doi.org/10.1037/0022-0663.90.1.37

Foorman, B. R., Francis, D. J., Winikates, D., Mehta, P., Schatschneider, C., & Fletcher, J. M. (1997). Early interventions for children with reading disabilities. *Scientific Studies of Reading, 1,* 255–276. http://dx.doi.org/10.1207/s1532799xssr0103_5

Frijters, J. C., Lovett, M. W., Sevcik, R. A., & Morris, R. D. (2013). Four methods of identifying change in the context of a multiple component reading intervention for struggling middle school readers. *Reading and Writing, 26,* 539–563. http://dx.doi.org/10.1007/s11145-012-9418-z

Frijters, J. C., Lovett, M. W., Steinbach, K. A., Wolf, M., Sevcik, R. A., & Morris, R. D. (2011). Neurocognitive predictors of reading outcomes for children with reading disabilities. *Journal of Learning Disabilities, 44,* 150–166. http://dx.doi.org/10.1177/0022219410391185

Fuchs, D., & Fuchs, L. S. (1998). Researchers and teachers working together to adapt instruction for diverse learners. *Learning Disabilities Research & Practice, 13,* 126–137.

Fuchs, D., & Fuchs, L. S. (2005). Peer-assisted learning strategies: Promoting word recognition, fluency, and reading comprehension in young children. *The Journal of Special Education, 39,* 34–44. http://dx.doi.org/10.1177/00224669050390010401

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Simmons, D. (1997). Peer-assisted learning strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34,* 174–206. http://dx.doi.org/10.3102/00028312034001174

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). *Reading First Impact Study Final Report Executive Summary* (Technical Report No. NCEE 2009-4039). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gardner, M. F. (1996). *Test of Visual Perceptual Skills (Non-Motor)—Revised*. San Fransisco, CA: Psychological and Educational Publications.

Gaskins, I. W., Downer, M. A., & Gaskins, R. W. (1986). *Introduction to the Benchmark School Word Identification/Vocabulary Development Program*. Media, PA: Benchmark School.

Gliner, J. A., & Morgan, G. A. (2000). *Research methods in applied settings: An integrated approach to design and analysis*. Mahwah, NJ: Erlbaum.

Glover, T. A., & Vaughn, S. (2010). *The Promise of response to intervention: Evaluating the current science and practice*. New York, NY: Guilford Press.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). *Does special education raise academic achievement for students with disabilities?* Paper presented at the National Bureau of Economic Research, Cambridge, MA. http://dx.doi.org/10.3386/w6690

Harris, A. D., McGregor, J. C., Perencevich, E. N., Furuno, J. P., Zhu, J., Peterson, D. E., & Finkelstein, J. (2006). The use and interpretation of quasi-experimental studies in medical informatics. *Journal of the Amer-*

ican Medical Informatics Association, 13, 16–23. http://dx.doi.org/10.1197/jamia.M1749

Hollingshead, A. B. (1975). *Four factor index of social status. Unpublished manuscript.* Yale University, New Haven, CT.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985031

Kamil, M. L. (2004). The current state of quantitative research. *Reading Research Quarterly, 39,* 100–108.

Kaufman, A. S., & Kaufman, A. L. (1990). *Kaufman Brief Intelligence Test.* Circle Pines, MN: American Guidance Service, Inc.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12,* 281–300. http://dx.doi.org/10.1080/10888430802132279

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22,* 354–367. http://dx.doi.org/10.1016/j.learninstruc.2012.02.001

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22,* 354–367. http://dx.doi.org/10.1016/j.learninstruc.2012.02.001

Kirby, J. R., Parrila, R., & Pfeiffer, S. (2003). Naming speed and phonological awareness as predictors of reading development. *Journal of Educational Psychology, 95,* 453–464. http://dx.doi.org/10.1037/0022-0663.95.3.453

Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology, 95,* 211–224. http://dx.doi.org/10.1037/0022-0663.95.2.211

Lovett, M. W., Barron, R. W., & Frijters, J. C. (2013). Word identification difficulties in children and adolescents with reading disabilities: Intervention research findings. In H. L. Swanson, K. Harris & S. Graham (Eds.), *Handbook of learning disabilities* (2nd ed., pp. 329–360). New York, NY: Guilford Press.

Lovett, M. W., Borden, S. L., DeLuca, T., Lacerenza, L., Benson, N. J., & Brackstone, D. (1994). Treating the core deficits of developmental dyslexia: Evidence of transfer-of-learning following phonologically- and strategy-based reading training programs. *Developmental Psychology, 30,* 805–822. http://dx.doi.org/10.1037/0012-1649.30.6.805

Lovett, M. W., Lacerenza, L., & Borden, S. L. (2000). Putting struggling readers on the PHAST track: A program to integrate phonological and strategy-based remedial reading instruction and maximize outcomes. *Journal of Learning Disabilities, 33,* 458–476. http://dx.doi.org/10.1177/002221940003300507

Lovett, M. W., Lacerenza, L., Borden, S. L., Frijters, J. C., Steinbach, K. A., & De Palma, M. (2000). Components of effective remediation for developmental reading disabilities: Combining phonological and strategy-based instruction to improve outcomes. *Journal of Educational Psychology, 92,* 263–283. http://dx.doi.org/10.1037/0022-0663.92.2.263

Lovett, M. W., Lacerenza, L., Steinbach, K. A., & De Palma, M. (2014). Empower™ Reading: Development and roll-out of a research-based intervention program for children with reading disabilities. *Perspectives on Language and Literacy* (Vol. 40, pp. 21–31). Baltimore, MD: International Dyslexia Association.

Lovett, M. W., Lacerenza, L., De Palma, M., & Frijters, J. C. (2012). Evaluating the efficacy of remediation for struggling readers in high school. *Journal of Learning Disabilities, 45,* 151–169.

Lovett, M. W., & Steinbach, K. A. (1997). The effectiveness of remedial programs for reading disabled children of different ages: Does the benefit decrease for older children? *Learning Disability Quarterly, 20,* 189–210. http://dx.doi.org/10.2307/1511308

Manis, F. R., Doi, L. M., & Bhadha, B. (2000). Naming speed, phonological awareness, and orthographic knowledge in second graders. *Journal of Learning Disabilities, 33,* 325–333, 374. http://dx.doi.org/10.1177/002221940003300405

Mason, L. H. (2004). Explicit self-regulated strategy development versus reciprocal questioning: Effects on expository reading comprehension among struggling readers. *Journal of Educational Psychology, 96,* 283–296. http://dx.doi.org/10.1037/0022-0663.96.2.283

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly, 40,* 148–182. http://dx.doi.org/10.1598/RRQ.40.2.2

Mathes, P. G., Howard, J. K., Allen, S. H., & Fuchs, D. (1998). Peer-assisted learning strategies for first-grade readers: Responding to the needs of diverse learners. *Reading Research Quarterly, 33,* 62–94. http://dx.doi.org/10.1598/RRQ.33.1.4

McMaster, K. N., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children, 71,* 445–463.

Morgan, P. L., Frisco, M., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education, 43,* 236–254. http://dx.doi.org/10.1177/0022466908323007

Morris, R. D., Lovett, M. W., Wolf, M., Sevcik, R. A., Steinbach, K. A., Frijters, J. C., & Shapiro, M. B. (2012). Multiple-component remediation for developmental reading disabilities: IQ, socioeconomic status, and race as factors in remedial outcome. *Journal of Learning Disabilities, 45,* 99–127. http://dx.doi.org/10.1177/0022219409355472

Morsy, L., Kieffer, M., & Snow, C. E. (2010). *Measure for measure: A critical consumers' guide to reading comprehension assessments for adolescents.* New York, NY: Carnegie Corporation of New York.

Nagy, W. E. (2007). Metalinguistic awareness and the vocabulary—Comprehension connection. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 52–77). New York, NY: Guilford Press.

Nakao, K., & Treas, J. (1992). *1989 Socioeconomic Index of Occupations: Construction from the 1989 Occupational Prestige Scores* (General Social Survey Methodological Report No. 74). Chicago, IL: National Opinion Research Center.

Nelson, J. R., Benner, G. J., & Gonzalez, J. (2005). An investigation of the effects of a prereading intervention on the early literacy skills of children at risk of emotional disturbance and reading problems. *Journal of Emotional and Behavioral Disorders, 13,* 3–12. http://dx.doi.org/10.1177/10634266050130010101

Newcomer, P. (1999). *Standardized Reading Inventory—2 (SRI-2).* Austin, TX: Pro-Ed.

O'Connor, R. E. (2000). Increasing the intensity of intervention in kindergarten and first grade. *Learning Disabilities Research & Practice, 15,* 43–54. http://dx.doi.org/10.1207/SLDRP1501_5

O'Connor, R. E., Fulmer, D., Harty, K. R., & Bell, K. M. (2005). Layers of reading intervention in kindergarten through third grade: Changes in teaching and student outcomes. *Journal of Learning Disabilities, 38,* 440–455. http://dx.doi.org/10.1177/00222194050380050701

OECD. (2013). *Education at a Glance 2013: OECD Indicators,* OECD Publishing. http://dx.doi.org/10.1787/eag-2013-en

Parrila, R., Kirby, J. R., & McQuarrie, L. (2004). Articulation rate, naming speed, verbal short-term memory, and phonological awareness: Longitudinal predictors of early reading development? *Scientific Studies of Reading, 8,* 3–26. http://dx.doi.org/10.1207/s1532799xssr0801_2

Pham, A. V., & Hasson, R. M. (2014). Verbal and visuospatial working memory as predictors of children's reading ability. *Archives of Clinical Neuropsychology, 29,* 467–477. http://dx.doi.org/10.1093/arclin/acu024

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials, 2,* 152–162. http://dx.doi.org/10.1191/1740774505cn076oa

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., & Torgesen, J. K. (2007). *Interventions for adolescent struggling readers: A meta-analysis with implications for practice.* Portsmouth, NH: Center on Instruction, RMC Research Corporation.

Scanlon, D. M., & Vellutino, F. R. (1997). A comparison of the instructional backgrounds and cognitive profiles of poor, average, and good readers who were initially identified as at risk for reading failure. *Scientific Studies of Reading, 1,* 191–215. http://dx.doi.org/10.1207/s1532799xssr0103_2

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, England: Sage.

Suggate, S. P. (2010). Why what we teach depends on when: Grade and reading intervention modality moderate effect size. *Developmental Psychology, 46,* 1556–1579. http://dx.doi.org/10.1037/a0020612

Swanson, H. L. (2000). Are working memory deficits in readers with learning disabilities hard to change? *Journal of Learning Disabilities, 33,* 551–566. http://dx.doi.org/10.1177/002221940003300604

Swanson, H. L. (2010). Does the dynamic testing of working memory predict growth in nonword fluency and vocabulary in children with reading disabilities. *Journal of Cognitive Education and Psychology, 9,* 51–77. http://dx.doi.org/10.1891/1945-8959.9.2.139

Swanson, H. L., Hoskyn, M., & Lee, C. (1999). *Interventions for students with learning disabilities: A meta-analysis of treatment outcomes.* New York, NY: Guilford Press.

Swanson, H. L., & Sáez, L. (2003). Memory difficulties in children and adults with learning disabilities. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 182–198). New York, NY: Guilford Press.

Swanson, H. L., Sáez, L., & Gerber, M. (2006). Growth in literacy and cognition in bilingual children at risk or not at risk for reading disabilities. *Journal of Educational Psychology, 98,* 247–264. http://dx.doi.org/10.1037/0022-0663.98.2.247

Swanson, H. L., & Siegel, L. S. (2001). Learning disabilities as a working memory deficit. *Issues in Education: Contributions from Educational Psychology, 7,* 1–48.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice, 15,* 55–64. http://dx.doi.org/10.1207/SLDRP1501_6

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34,* 33–58, 78. http://dx.doi.org/10.1177/002221940103400104

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1997a). Approaches to the prevention and remediation of phonologically-based reading disabilities. In B. A. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 287–304). Mahwah, NJ: Erlbaum.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1997b). Prevention and remediation of severe reading disabilities: Keeping the end in mind. *Scientific Studies of Reading, 1,* 217–234. http://dx.doi.org/10.1207/s1532799xssr0103_3

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency (TOWRE).* Austin, TX: Pro-Ed.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Alexander, A. W., & Conway, T. (1997). Preventive and remedial interventions for children with severe reading disabilities. *Learning Disabilities: A Multidisciplinary Journal, 8,* 51–62.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91,* 579–593. http://dx.doi.org/10.1037/0022-0663.91.4.579

Vadasy, P. F., Sanders, E. A., Peyton, J. A., & Jenkins, J. R. (2002). Timing and intensity of tutoring: A closer look at the conditions for effective early literacy tutoring. *Learning Disabilities Research & Practice, 17,* 227–241. http://dx.doi.org/10.1111/1540-5826.00048

Vaughn, S., Chard, D. J., Pedrotty-Bryant, D., Coleman, M., Tyler, B.-J., Linan-Thompson, S., & Kouzekanani, K. (2000). Fluency and comprehension interventions for third-grade students. *Remedial and Special Education, 21,* 325–335. http://dx.doi.org/10.1177/074193250002100602

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice, 18,* 137–146. http://dx.doi.org/10.1111/1540-5826.00070

Vaughn, S., Levy, S., Coleman, M., & Bos, C. S. (2002). Reading instruction for students with LD and EBD: A synthesis of observation studies. *The Journal of Special Education, 36,* 2–13. http://dx.doi.org/10.1177/00224669020360010101

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69,* 391–409. http://dx.doi.org/10.1177/001440290306900401

Vaughn, S., Wexler, J., Roberts, G., Barth, A. A., Cirino, P. T., Romain, M. A., . . . Denton, C. A. (2011). Effects of individualized and standardized interventions on middle school students with reading disabilities. *Exceptional Children, 77,* 391–407.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88,* 601–638. http://dx.doi.org/10.1037/0022-0663.88.4.601

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing.* Austin, TX: Pro-Ed.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review, 36,* 541–561.

Wechsler, D. (1997). *Wechsler Intelligence Scale for Children* (3rd ed.) San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.) San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence (WASI).* New York, NY: The Psychological Corporation.

Wiederholt, J. L., & Bryant, B. R. (2004). *Gray Oral Reading Tests* (4th ed.). Austin, TX: Pro-Ed.

Wilkinson, G. S. (1993). *Wide Range Achievement Test—3.* Wilmington, DE: Jastak.

Wolf, M., Barzillai, M., Gottwald, S., Miller, L., Spencer, K., Norton, E., . . . Morris, R. (2009). The RAVE-O intervention: Connecting neuroscience to the classroom. *Mind, Brain and Education: 3,* 84–93. http://dx.doi.org/10.1111/j.1751-228X.2009.01058.x

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology, 91,* 415–438. http://dx.doi.org/10.1037//0022-0663.91.3.415

Wolf, M. A., & Denckla, M. (2005). *The rapid automatized naming and rapid alternating stimulus tests.* Austin, TX: Pro-Ed.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading, 5,* 211–239. http://dx.doi.org/10.1207/S1532799XSSR0503_2

Wolf, M., Miller, L., & Donnelly, K. (2000). Retrieval, Automaticity, Vocabulary Elaboration, Orthography (RAVE-O): A comprehensive, fluency-based reading intervention program. *Journal of Learning Disabilities, 33,* 375–386. http://dx.doi.org/10.1177/002221940003300408

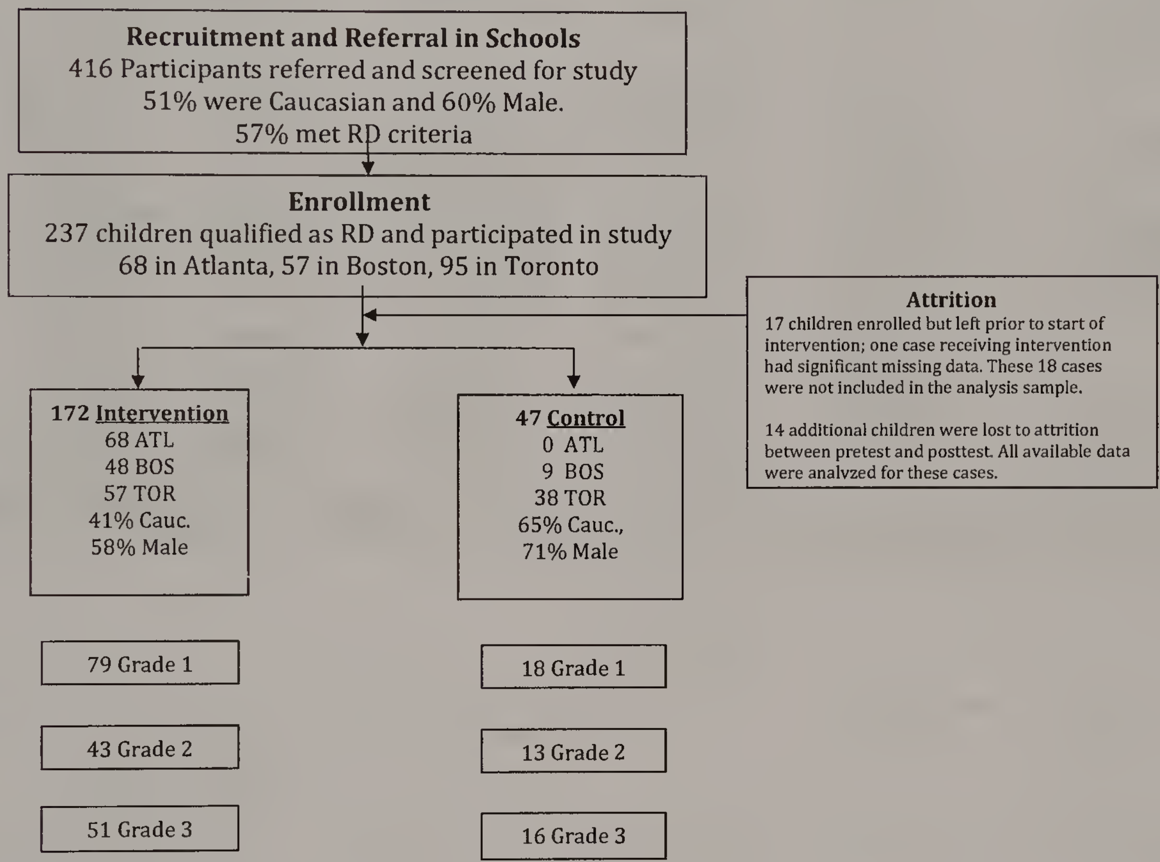
Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests—Revised.* Circle Pines, MN: American Guidance Service.

Appendix

Descriptive Details About Research Procedures, Sub-Sample Profiles, and the Triple-Focus Intervention Lessons

Table A1

Flowchart Illustrating Recruitment, Assignment, and Intervention for First, Second, and Third Grade Participants in Atlanta, Boston, and Toronto



(Appendix continues)

Table A2
Descriptive Statistics for Every Outcome Measure, Subdivided by Measurement Occasion, Intervention Condition, and Grade

Measure	Intervention			Control		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
Pretest						
Sound combinations	1.71 (2.30)	5.91 (4.50)	10.37 (3.98)	2.33 (3.09)	4.58 (4.32)	11.08 (3.55)
Challenge Test	0.01 (0.12)	1.44 (3.03)	13.30 (12.76)	0.00 (0.00)	0.18 (0.60)	11.31 (11.19)
TOWRE Phonological Decoding	1.14 (1.93)	3.14 (3.44)	8.20 (5.22)	0.71 (1.33)	2.08 (3.45)	7.17 (4.32)
TOWRE Sight Words	4.42 (4.20)	18.21 (10.62)	35.75 (12.03)	5.33 (6.31)	14.85 (8.07)	31.00 (11.52)
WRAT-3 Reading	14.61 (2.29)	18.91 (3.39)	23.45 (3.09)	13.72 (3.53)	17.54 (3.23)	23.25 (2.93)
WRMT-R Word Attack	438.28 (7.46)	450.00 (13.65)	466.04 (12.74)	436.44 (6.13)	443.85 (10.80)	468.88 (10.07)
WRMT-R Word Identification	358.90 (15.63)	402.84 (24.51)	440.00 (16.41)	360.00 (18.67)	394.69 (21.48)	428.56 (25.17)
WRMT-R Passage Comp.	412.71 (10.78)	435.7 (17.84)	461.06 (12.45)	413.28 (11.58)	426.69 (15.34)	459.63 (17.81)
GORT-R Comprehension	1.34 (1.81)	3.21 (3.28)	12.20 (8.21)	1.20 (1.78)	4.08 (3.77)	12.31 (8.69)
SRI Comprehension	0.45 (0.64)	6.16 (7.98)	16.71 (10.58)	1.15 (3.29)	4.00 (5.79)	19.42 (11.84)
GORT-R rate	0.34 (0.67)	2.40 (3.09)	9.29 (5.24)	0.33 (0.62)	2.23 (2.95)	7.25 (4.80)
Multiple definitions trained	0.81 (0.29)	0.90 (0.37)	1.24 (0.26)	0.82 (0.33)	0.88 (0.23)	1.19 (0.23)
WORD-2	18.99 (5.04)	29.28 (7.98)	39.96 (10.09)	18.44 (4.03)	27.69 (4.94)	37.38 (9.65)
PIAT-R Spelling	1.71 (2.30)	5.91 (4.50)	10.37 (3.98)	2.33 (3.09)	4.58 (4.32)	11.08 (3.55)
Posttest						
Sound combinations	14.75 (5.27)	18.24 (6.35)	20.20 (4.67)	7.12 (6.36)	6.75 (3.17)	11.56 (5.23)
Challenge Test	15.44 (13.58)	25.71 (14.20)	34.84 (9.43)	4.88 (10.20)	.58 (1.00)	17.75 (11.65)
TOWRE Phon. Decoding	10.75 (7.05)	11.05 (7.42)	15.71 (6.77)	6.00 (7.16)	2.85 (3.05)	8.13 (5.03)
TOWRE Sight Words	28.49 (12.78)	34.4 (11.86)	47.94 (12.17)	17.13 (11.37)	22.38 (8.91)	37.81 (12.53)
WRAT-3 Reading	24.27 (4.18)	25.79 (3.57)	28.76 (2.61)	19.53 (5.14)	19.85 (2.70)	25.06 (3.66)
WRMT-R Word Attack	474.86 (12.92)	476.88 (11.45)	481.39 (10.38)	455 (17.86)	454.85 (10.33)	469.75 (9.50)
WRMT-R Word Identification	429.73 (20.83)	442.33 (18.79)	461.25 (13.84)	401.28 (30.54)	415.69 (20.55)	441.19 (23.12)
WRMT-R Passage Comp.	453.33 (14.96)	465.86 (14.6)	479.94 (8.13)	435.47 (17.39)	443.00 (16.05)	466.38 (12.98)
GORT-R Comprehension	8.71 (8.06)	15.8 (9.17)	22.28 (8.46)	5.59 (6.19)	6.77 (6.88)	16.13 (10.16)
SRI Comprehension	12.48 (9.51)	19.24 (11.28)	27.9 (10.54)	6.12 (7.36)	7.38 (7.69)	22.44 (11.03)
GORT-R Rate	6.67 (5.57)	9.46 (5.84)	15.59 (6.31)	3.76 (4.01)	4.46 (3.18)	9.8 (5.99)
Multiple Definitions Trained	1.27 (.35)	1.56 (.45)	1.80 (.35)	.97 (.29)	.86 (.23)	1.19 (.31)
WORD-2	34.65 (7.53)	41.5 (6.92)	51.37 (10.45)	26.89 (8.30)	32.85 (6.00)	39.80 (8.30)
PIAT-R Spelling	14.75 (5.27)	18.24 (6.35)	20.2 (4.67)	7.12 (6.36)	6.75 (3.17)	11.56 (5.23)

Note. WRMT-R = Woodcock Reading Mastery Test—Revised; SRI = Standardized Reading Inventory Comprehension; TOWRE = Test of Word Reading Efficiency; GORT-R = Gray Oral Reading Test–Version 4; SRI = Standardized Reading Inventory Comprehension; WORD-2 = The WORD Test 2—Elementary; PIAT-R = Peabody Individual Achievement Test—Revised.

(*Appendix continues*)

Table A3
The Triple-Focus Reading Program: An Overview of Lessons 32, 77, and 106

Instructional Components (time)	Lesson 32	Instructional Components (time)	Lesson 77	Instructional Components (time)	Lesson 106
Metacognition (3–5 min)	• Review program goals (e.g., *Why is it important to learn how to read? What do you want to be able to read?*) • Strategy Skills Review #2 (Metacognitive Dialogue for decoding strategies; e.g., *How many strategies have we learned? When/How/Why do we use the strategy?*)	***Metacognition (3–5 min)***	• Strategy Skills Review #4 (Metacognitive Dialogue for all strategies) • Game Plan review (Metacognitive dialogue selection and application of multiple strategies; e.g., *Which strategy(ies) would you choose to figure out this word? Why?*)	***Metacognition (3–5 min)***	Comprehension Strategy Skill Review (Metacognitive Dialogue to review comprehension strategies (e.g., *What strategy(ies) are we going to use to understand text? What is the first thing we do to clarify? Whot ore the 4 questions we osk ourselves after reading the beginning of a story?*)
Sounding Out (10 min)	New sound: **j**; Reading Vocabulary: *shop, soon, must, never, hop, tolked, was, day, walked, brush, brushed;* Workbook activities	***Sounding Out (10 min)***	Reading Vocabulary: *nothing, bine, bin, flower, anyhow, magic, scream, soying, holding, picked, spells, you're, biggest, can't, salt, doesn't . . . side, hopper, hoper, fine, fin;* Story: *The Ghosts Turn on Boo*	***Sounding Out (5 min)***	**Reading Vocabulary: soft, third, first, cried, tried, could, thud, mountain, eaten, cloud, chop, shouted, eating, around, stepped, grape, bananas, grabbed, . . . disappear, striped, paths, myself, more;** Story: *Jean Eats Red Banonos*
Rhyming Strategy (15 min)	Word identification by Analogy. • Teacher models Rhyming Strategy and introduces new keywords • Students complete worksheet (e.g., *If I know "luck", then I know "truck...)*	***Rhyming Strategy (10 min)***	Word identification by Analogy. • Teacher models Rhyming Strategy and introduces new keywords • Students complete worksheet (e.g., *If I know "page", then I know "stage.")*	***Comprehension Instruction (25–30 min)***	Metacognitive comprehension strategies (predicting, summarizing, clarifying, questioning) to improve student's engagement with and understanding of text.
Words of week	***bug, luck, bus***	Words of week	***page, boat, food, fool***	Application to text	The Plot Graph; applying the Questioning Strategy to Fiction (e.g., *At the beginning of a story, we ask 4 questions to focus our attention on the important characters and events: the 4Ws: Who? When/Where does the story take ploce? What is the problem?)*
Teacher Model Sentence	The police took a **mugshot** of us in the **truck.**	Teacher Model Sentence	**On stage,** the **goat** was in a bad **mood** and kicked the **stool.**		
Student Model Sentence	The **bug** ran out of **luck.**	***Vocabulary Development (10 min)***	Activities to develop vocabulary breadth and depth (e.g., understanding homonyms, building word meanings, exploring impact of prefixes/suffixes).		
Vocabulary Development (5 min)	Activities to develop vocabulary breadth and depth (e.g., understanding homonyns, building word meanings, exploring impact of prefixes/suffixes).	Harder Starters	***squ***	***Vowel Alert Strategy (5–10 min)***	Flexibility with the variant sounds of vowels, vowel combinations, and other vowel concepts (e.g., –ol, –al, ul).
Multiple meaning words[a]	***bug, luck*** (e.g., *Whot does bug meon? What would someone meon if he says, "This ploce bugs me." If a bug londs on your arm, what do you do?)*	Word Web[a]	***place*** (A semantic activity that provides a visual way of illustrating how words are interconnected and gives visual images to aid memory.)	Vowel/concepts	Review L-spells (i.e., l-controlled vowels –al, –ol); introduce –ul (lesson + worksheet)
Peeling Off Strategy (10 min)	Identification of affixes in multisyllabic words (e.g., *I peel off ___ at the end of the word. The root is ___. The word is ___.)*	***Peeling Off Strategy (5–10 min)***	Identification of affixes in multisyllabic words *I peel off___ at the beginning/end of the word. The root is ___. The word is ___.*	***Vocabulary Development (10 min)***	Activities to develop vocabulary breadth and depth (e. g., understanding homonyns, building word meanings, exploring impact of prefixes/suffixes).

(Appendix continues)

Table A3 (*continued*)

Instructional Components (time)	Lesson 32	Instructional Components (time)	Lesson 77	Instructional Components (time)	Lesson 106
Affixes	*–ed* (past tense); *–er* (one who) (lesson + worksheet)	Affixes	*di–* (lesson + worksheet)	Multiple meaning words[a]	***bug, luck*** (e.g., *What does bug mean? What would someone mean if he says, "This place bugs me." If a bug lands on your arm, what do you do?)*
Application Activities[b] *(15 min)*	• Speed Wizard (computer activity to develop automaticity and fluency) • Minute Story (application of strategies and skills to text)	***Vowel Alert Strategy (5–10 min)***	Flexibility with the variant sounds of vowels and combinations. (e.g., *I see the double trouble twin –**ow** and* underline it with two lines. First I try *–**ow*** as in glow; then I try **–ow as in cow. I go when I get a word that makes sense.)**	***Peeling Off Strategy OR SPY Strategy (5–10 min)***	Alternate activities that (i) review, consolidate, and apply all affixes introduced with activities appropriate to the application of the SPY Strategy.
		Vowels	Review vowel combinations: ***–oo, –ea, –ow, –ie***		
		SPY Strategy (5 min)	Finding small words in larger words (e.g., *I SPY "base" I SPY "ball." The word is "baseball."*	*Application Activities*[b] *(5 min)*	Focus is on consolidation and automaticity; alternate activities such as Speed Wizard and Challenge Words Challenge words
		Application Activities[b] *(5 min)*	Challenge words (multisyllabic words on which the students can apply the decoding strategies)		

[a] Each week, 2–4 multiple meaning words were introduced and 1 Word Web was completed. [b] Application Activities alternated each day throughout program.

Received August 28, 2015
Revision received November 22, 2016
Accepted November 23, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 7, 915–934

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000188

Streaming, Tracking and Reading Achievement: A Multilevel Analysis of Students in 40 Countries

Ming Ming Chiu
Purdue University

Bonnie Wing-Yin Chow
City University of Hong Kong

Sung Wook Joh
Seoul National University

Grouping similar students together within schools (*streaming*) or classrooms (*tracking*) based on past literacy skills (reported by parents), family socioeconomic status (SES) or reading attitudes might affect their reading achievement. Our multilevel analysis of the reading tests of 208,057 fourth-grade students across 40 countries, and their parents', teachers', principals', and their survey responses yielded the following results. Streaming was linked to higher reading achievement (consistent with *customized instruction*), but tracking was linked to lower reading achievement (consistent with more *help opportunities*). Students had higher reading achievement when classmates had stronger past literacy skills (suggesting *sharing ideas*) or extremely poor ones (help opportunities). Also, when classmates have higher family SES, students had higher reading achievement (suggesting *sharing resources*), with diminishing marginal returns. When classmates' family SES differed more (more *diversity*), students with greater past literacy skills had higher reading achievement (*Matthew effect*). Lastly, when classmates had better reading attitudes, students with lower past literacy skills showed higher reading achievement (*modeling, norms*). When classmates' reading attitudes differed more, students had higher reading achievement (*contrasting cases*), although extreme differences weakened this link (less *homophily*). These results suggest that streaming across schools and mixing of students within classrooms (by past achievement, family SES and reading attitudes) are linked to overall reading achievement.

Keywords: ability grouping, classmates, inequality, international assessment, socioeconomic status

Grouping students by ability across classrooms (*tracking*) is a common but controversial education policy. Past studies of tracking have yielded mixed results (*positive effects*: e.g., Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Chmielewski, Dumont, & Trautwein, 2013; *negative*: e.g., Guyon, Maurin, & McNally, 2012; Hanushek & Wößmann, 2006; Jakubowski, Patrinos, Porta, & Wiśniewski, 2016; and *nonsignificant*: e.g., Opdenakker & Van Damme, 2001). To explain these contradictory results, we propose that the impact of ability grouping on student achievement differs across levels (*school streaming* vs. *classroom tracking*), differs across types of classmate resources (past achievement, attitude, family socioeconomic status [SES]), and depends on a student's own academic ability.

This article was published Online First March 6, 2017.

Ming Ming Chiu, Department of Educational Studies, Purdue University; Bonnie Wing-Yin Chow, Department of Applied Social Sciences, City University of Hong Kong; Sung Wook Joh, Department of Finance, Seoul National University.

We appreciate the financial support from the Institute of Management Research at Seoul National University.

We appreciate the research assistance of Yik Ting Choi.

Correspondence concerning this article should be addressed to Ming Ming Chiu, Department of Educational Studies, Purdue University, 5156 Beering Hall, 100 North Street, West Lafayette, IN 47907. E-mail: chiu23@purdue.edu

Data or analytic limitations might also account for some conflicting results. While some education systems stream or track openly, others do so quietly without a public policy, even though students can often recognize different achievement patterns across groups (Alexander, Entwisle, & Dauber, 2003; Boaler, 2013; Lucas & Berends, 2002). Hence, we examine the distribution of students by their actual past achievement, instead of deferring to incomplete public declarations of streaming or tracking policies (Ding & Lehrer, 2007; Hanushek, Kain, Markman, & Rivkin, 2003). Analytic limitations of past studies include omitted variable bias, multicollinearity, and failure to model nested data structure of students within schools within countries (Goldstein, 2011; Kennedy, 2008).

To disentangle the effects of different types of streaming and tracking on academic achievement, this study considers different levels of factors and extends past research in four ways via multilevel analyses of 208,057 fourth grade primary school students in 40 countries. First, we examine whether school streaming or classroom tracking is related to reading achievement. Second, we examine whether the amount of or variation in different types of classmate resources (past achievement, reading attitude, family SES) are related to a student's reading achievement. Third, we analyze whether these relations differ across students with different levels of reading achievement (bottom 10%, 20%, and 50% and top 50%, 20% and 10% from each country). Lastly, we overcome the statistical limitations of past studies through repre-

sentative sampling, inclusion of central variables, structured sets of variables, and multilevel analyses. By understanding how streaming and tracking operate, this study can help explain different effects and inform government and school policies regarding assignment of students to schools and classrooms to improve student learning.

Student Grouping by Ability and Student Achievement

Many education systems group students with similar past academic achievements together for instruction. Whether this approach raises or reduces student achievement depends on (a) the impact of student similarities versus differences and (b) whether classmates compete or share resources.

Impact of Student Similarities Versus Differences

Education administrators can place students with similar academic competences together, separating them from other students with higher or lower levels of past academic achievement. Or, they may mix together students with different past achievements.

Clusters of similar students. Students with similar past academic achievement can be assigned together to the same school (*academic* or *vocational streaming*, Chmielewski, 2014; also known as *banding*, Chiu & Walker, 2007) or to the same classes within a school (*course-by-course tracking*, Chmielewski, 2014). Grouping similar students together can improve their learning by enabling teachers to customize instruction for similar students (Watanabe, 2008), capitalizing on student preferences to interact with and help similar peers (Brechwald & Prinstein, 2011), or enacting self-fulfilling prophecies of their labels and norms (Kaplan, Guzman, & Tomlinson, 2009: Pintrich, 2003).

When facing students with similar academic competences, educators can customize the curriculum, lessons, teaching materials, and teaching pace to the needs of each group of similar students, which can improve their learning (*customized instruction*, Smith, 2013). In contrast, when the academic competences of students differ widely, teachers may focus on the learning needs of a subset of students to the detriment of other students with much higher or much lower academic competences (Westwood, 2013). Implementing customized instruction is easier for a streamed school with similar students than for only a class with similar students (but different students across classes; Watanabe, 2008). In a streamed school, teachers design lessons for all students within a small range of competences (Chiu & Walker, 2007). In contrast, classroom tracking requires much more teacher time and effort to create different lessons for classes of students with different competences (Smith, 2013). Hence, we expect more instruction customization and higher student achievement in streamed schools than in other schools (with or without classroom tracking).

As students often prefer to interact with others who are similar to themselves, those with similar academic competences might be more likely than others to interact, befriend, and help one another to learn, compared to dissimilar others (*homophily*, Brechwald & Prinstein, 2011; aka *assortativeness*, Kindermann, 2007). As a result, academically similar students in streamed schools might be more likely to help one another and contribute to a school-wide community culture of mutual support, compared to students in nonstreamed schools (Chiu, 2008). (While academically similar classmates in tracked classes within a nonstreamed school might help one another more than less similar classmates would, they might be less helpful to academically different *schoolmates* in other classes.) As the differences between two students increase, their likelihood of interacting and helping one another decreases (Brechwald & Prinstein, 2011).

Although students placed in schools or classes labeled as high-achieving might benefit from self-fulfilling prophecies by enacting high expectations and norms (*assimilation,* Jansen, Schroeders, Lüdtke, & Marsh, 2015), students labeled as low-achieving might correspondingly suffer. Teachers and parents of students in schools labeled as high-achieving typically have high expectations of them, which students often internalize (Pintrich, 2003). As a result, these students, their teachers, and their parents tend to invest more time, effort, and other resources to improve their learning compared to those in nonstreamed schools (Kaplan et al., 2009).

Assimilation effects might be stronger when tracking across classrooms rather than streaming across schools. When students are tracked across classrooms within a school, teachers can devote more attention, effort, and other resources to students in higher tracks than to students in lower tracks (Chiu & Khoo, 2005). This unequal distribution of teacher resources not only increases the gap between high- and low-achieving students, but its unfairness can demoralize low-achieving students (Chiu, 2008). As a result, the drop in the academic performance of low-achieving students might exceed the gain in that of high-achieving students, thereby yielding lower overall academic performance (Chiu, 2008).

In contrast, the negative effects of assimilation and demoralization might be weaker for students within a streamed school. As academic comparisons with other schools are more distant from students and teachers' immediate experiences, such labels might have less impact on their behaviors, especially as they acclimate to their streamed school environment (Chiu, 2008). (Such labels can still influence parents and attract teachers to reputable schools in education systems with open markets for hiring teachers [unlike school systems like South Korea that randomly rotate teachers to different schools every five years, Chiu & Khoo, 2005].) Furthermore, students within a streamed school share the same label, so teachers and staff are less likely to treat students differently (Chiu & Khoo, 2005). As a result, these students in streamed schools might be more likely to view their teachers as fair, to be less demoralized, and show higher overall academic performance compared to students in schools with tracked classes.

Hence, streamed schools might have more benefits from customized instruction and homophily, and less harm from assimilation and demoralization, compared to tracked classes. As many education systems do not announce an official streaming policy, we operationalized the degree of streaming across schools via *school clustering* by past achievement measure (ratio of student past literacy skills variance across schools over total variance of past literacy skills in a country [Chiu & Khoo, 2005]). Hence, we propose the following hypothesis:

> *H-1a:* In education systems with greater school clustering by past achievement, students have higher academic achievement than otherwise.

Mixing different students. While clustering similar ability students together aids instruction customization and capitalizes on

homophily, mixing different ability students together can facilitate helping behaviors and learning from schoolmates/classmates with different experiences (Chiu & Khoo, 2005). Mixed classes offer more possible pairs of a higher-achieving student helping a lower-achieving student (*help opportunities*, Chiu & Chow, 2015). Helping benefits both the recipient who receives additional information and explanation, and the giver who often learns more by reorganizing and elaborating his or her knowledge to give a suitable explanation to the recipient (Blatchford, Pellegrini, & Baines, 2016). In extremely streamed schools, however, students have similar achievement levels, so they have fewer opportunities to give or receive help, and consequently less learning (Chiu, 2008).

> *H-1b:* In education systems extremely clustered by past achievement, the positive impact on academic achievement is smaller.

Within a classroom, greater variance in past achievement among students yields more pairs of students with different past achievement levels, more opportunities to give or receive help, and possibly more learning (Chiu, 2008).

> *H-2a:* When classmates have greater *variance* in their past achievements, a student has higher academic achievement.

However, extreme differences in past achievement levels of students in a classroom can hinder high-achieving students' efforts to help lower achieving students. Students with higher past achievement are more likely to give help than receive help, so like teachers, they likely have more difficulty helping weaker students with extremely different achievement levels than those with similar achievement levels (*customized instruction*, Smith, 2013); as a result, they might learn less from such challenging (and potentially frustrating) help opportunities.

> *H-2b:* In classrooms with extremely high variance of past achievement among students, the positive impact on academic achievement is smaller, especially for higher-achieving students.

Compared to less diverse groups, more diverse groups often have weaker interpersonal relations, more disagreements, and less early learning, but their greater range of experiences and resolution of disagreements can increase their later learning (Watson, Johnson, & Zgourides, 2002). Due to *homophily bias* within a group, members often categorize themselves into subgroups based on similarity (similar to ingroup members and different from outgroup members, *social categorization*, van Knippenburg & Schippers, 2007). People trust ingroup members more, cooperate with them more often, and have better relationships with them, compared to outgroup members. Moreover, diverse groupmates often have unfamiliar ideas, attitudes, and experiences (Sharan, 2010) that can conflict, thereby igniting disagreements that hinder interpersonal relations initially (van Knippenberg, De Dreu, & Homan, 2004). Thus, less diverse groups often initially function better than more diverse groups do (Watson et al., 2002).

However, diverse groups' different ideas and disagreements can legitimize different opinions, thereby stimulating group members to pay attention, share more ideas, and reduce premature consensus (De Dreu & West, 2001). Group members' diverse views also help them recognize flaws and correct them to yield better ideas (Chiu, 2008). If diverse groupmates can reconcile their different views, resolve their disagreements, and integrate their information, they can create and learn new ideas (Paulus & Brown, 2003). Lastly, divergent views can stimulate a group to reflect and improve on its own functioning (Schippers, Den Hartog, & Koopman, 2007). Thus, over longer time periods, diverse groups produce ideas that are more diverse and learn more compared to homogeneous groups (van Offenbeek, 2001; Watson et al., 2002). As students with higher past achievement have more knowledge than other students, they might capitalize on it to integrate it with these new, different ideas more effectively (*Matthew effect*, Rigney, 2013). As the students in these data have shared a classroom for several months, we focus on the long-term effects of diversity. Also, many countries do not have much racial diversity across classmates, so we examine diversity of classmates' family SES and propose the following hypothesis.

> *H-3:* When classmates have greater *variance* in their family SES, a student has higher academic achievement, especially benefiting students with higher past achievement.

Greater differences among classmates increase the extremes of diametric opposites, which can draw attention to them and aid learning. Consider two classes whose students' reading attitudes have the same mean but greater variance in the second class than the first. As reading attitude extremes are likely greater in the second class than the first (Kennedy, 2008), the second class likely has both the student with the best reading attitude (let us call her Heidi) and the one with the worst (Lola). Heidi's concrete behaviors embody her reading attitude (e.g., reads many books; tells their stories by acting them) and yield beneficial consequences (high reading quiz scores; smiles at her graded quizzes; teacher praises her; reading awards; etc.). In contrast, Lola rarely reads ("I never read books—they're boring"), has low reading quiz scores, frowns at her graded quizzes, and so on.

As extremes, Heidi and Lola mutually highlight their differences (*contrasting cases*) and draw students' attention to their differences (*focusing function*, Schwartz & Bransford, 1998). This focusing function helps classmates recognize important differences between Heidi and Lola's reading attitudes and consequences (Roelle & Berthold, 2015). As contrasting cases of people rather than abstract ideas, Heidi and Lola serve as detailed reference points for inferences about students in the continuum between them (*cognitive reference point reasoning*, Tribushinina, 2008). The instantiation of positive reading attitudes and consequences in Heidi and their negative counterparts in Lola helps students organize two linked sets of reading attitudes and consequences, thereby learning effectively and efficiently about reading attitude, remembering it reliably, and acting on it accordingly to learn more than otherwise (Glogger et al., 2013). Hence, we propose this hypothesis:

> *H-4a:* When classmates have greater *variance* in their reading attitudes, a student has higher reading achievement.

As noted above, extreme differences in students in a classroom can sharply reduce their inclinations to interact and help one another (less homophily), which can reduce overall student learning (Chiu, 2008). Less homophily might also apply to extreme differences in reading attitudes.

H-4b: In classrooms with extremely high variance of reading attitude among students, the positive impact on academic achievement is smaller.

Classmates Compete Versus Share Resources

The impact of classroom tracking on student achievement also depends on the extent to which classmates compete or share resources. When competing with classmates with greater cognitive, social, and material resources in a zero-sum game, a student could have lower academic achievement. Classmates can serve as a collective ruler against which to judge a student's relative competence (*comparative reference-group view*). When surrounded by lower-achieving classmates, a student often has greater confidence in his or her competence (self-concept), expectation of future success, motivation, and academic achievement (*social comparison*, Liu, Wang, & Parkins, 2005). Conversely, higher-achieving classmates can demoralize a student, reduce his or her self-concept, lower future expectations, and yield lower academic achievement.

H-5a: When classmates have higher *mean* past academic achievement, a student has lower academic achievement.

On the other hand, classmates, especially high-achieving ones, might help a student learn directly or indirectly (Skibbe, Phillips, Day, Brophy-Herb, & Connor, 2012). Classmates can directly help a student by sharing information, which the student can use to learn more than otherwise (Chiu & Chow, 2015). For example, a higher-achieving classmate can help a student correctly spell an unfamiliar word. Alternate H-5b competes with H-5a above:

H-5b: When classmates have higher *mean* past academic achievement, a student has higher academic achievement.

As noted above, helping opportunities provide a competitor to both Hypotheses 5a and 5b, namely that more classmates whose past achievements differ from a student offer more help opportunities, either as givers or recipients of help (Chiu & Chow, 2015). Hence, when classmates' *mean* past academic achievement is much higher or much lower than a student, he or she has more help opportunities and learns more. Note that this hypothesis is conceptually equivalent to H-2a but is measured and tested through different variables; specifically, the results have (a) separate variables for *mean* past academic achievement and *variance* of past academic achievement, and (b) separate results for the subsamples of the bottom 10%, 20% and 50% and the top 50%, 20% and 10% of students.

H-5c: When the *mean* past academic achievement of classmates differs more from that of a student, he or she has higher academic achievement.

A classmate can also help students learn indirectly through motivation or norms. For example, a classmate can dramatically enact a scene from a storybook, which can entice and motivate other students to discuss it and learn about it (Edmunds & Bauserman, 2006; Guthrie, Klauda, & Morrison, 2012; Skibbe et al., 2012). Over time, students' greater motivation helps them exert more effort and persevere when facing difficulties (Chiu & Chow, 2010).

Classmates, especially higher-achieving ones with higher status, can help model, create and maintain norms of positive attitudes toward reading, regular learning behaviors, and high academic achievement (Chiu & Chow, 2015). Classmates can articulate and model positive academic attitudes, such as sharing their enjoyment of specific stories. Furthermore, they can discuss their readings daily to promote peer pressure toward regular reading of new books. Together, classmates can cultivate a culture of positive reading attitudes and behaviors in which to immerse a student, which typically yields higher academic achievement (Chiu & McBride-Chang, 2006; Johnson & Johnson, 1999). Students with low past achievement are less likely than those with high past achievement to have positive attitudes toward reading, so they are more likely to benefit from classmates with better attitudes toward reading.

H-6: Among students with lower past reading achievement, they have higher reading achievement when their classmates have higher *mean* attitude toward reading.

Students can benefit not only from their own family resources but also from classmates' family resources. Families can use their financial, human, cultural, and social capital to give their children learning opportunities. Specifically, families with more money (*financial capital*) can buy more educational resources (books, calculator, computer, etc.) to create a richer learning environment (D. P. Baker, Goesling, & Letendre, 2002). Furthermore, students in high SES families can benefit more from their parents' human, social, and cultural capital. Families with more education, knowledge, or skills (*human capital*) often create better learning environments for their children, foster better attitudes toward reading, and teach them more skills compared to other families (Davalos, Chavez, & Guardiola, 2005; Willms, 1999). Likewise, high SES families often have cultural possessions or experiences (*cultural capital*) that can help their children learn society's cultural knowledge, skills, and values to adapt to their school culture (Lee & Bowen, 2006). High SES families also often have large social networks of relatives, friends, and acquaintances with skills or resources (*social capital*) that can help their children learn (Israel, Beaulieu, & Hartless, 2001). Using their greater financial, human, social, and cultural capital, higher SES students can better understand others' expectations, behave properly at school, have closer relationships with teachers and classmates, and learn more both at home and in school than lower SES students do.

Similarly, a student can benefit from a classmate's family resources directly or indirectly (*spillover externalities*; Chiu & Chow, 2015; Mankiw, 2014). In the most direct case, a student visits a classmate's home and uses the latter's educational resources. A student may work with a classmate on the latter's computer (classmate family financial capital), discuss a book with the classmate's mom (classmate family human capital), discuss a painting in the living room (classmate family cultural capital), or chat with a family friend over dinner (classmate family social capital). Less directly, a student can learn from a classmate's learning experiences at home (Chiu & Chow, 2015) when the classmate talks about it ("my mom was telling me about the presidential election").

Furthermore, classmate family SES and a student's academic achievement might have a nonlinear relation. Consider, for exam-

ple, a student with three calculators. The first calculator is extremely useful for fast and accurate arithmetic computations. The second calculator might be marginally helpful as a backup, but is less useful than the first one. A third calculator offers little additional benefit. The high value of the first calculator, the lower value of the second one, and the miniscule value of the third one is an example of *diminishing marginal returns* in which each additional object has less value, ceteris paribus. Past studies suggest that resources often have diminishing marginal returns indicated by a logarithmic relation (e.g., Chiu, 2015).

> *H-7:* When classmates have a higher *mean* family SES, a student has higher academic achievement, although with diminishing marginal returns.

The Present Study

This study examines the effects of school streaming and classroom tracking on students' reading achievement through analyses of 208,057 fourth grade primary school students in 40 countries. (Many primary schools track their students, some as early as first grade [Alexander et al., 2003; Ansalone & Biafora, 2004; Lleras & Rangel, 2009].) We focus on three research questions. First, is school streaming or classroom tracking linked to higher reading achievement? Second, do these links operate through the mechanisms of *customized instruction, labeling, homophily, help opportunities, diversity, Matthew effect, contrasting cases, social comparison, sharing resources*, or *modeling/norms*? Lastly, do these relations differ across subsamples of the lowest-achieving 10%, 20%, and 50% of students and the highest-achieving 50%, 20% and 10% of students in each country?

As past studies have shown that reading achievement is related to the following variables, we included them in our regression model to reduce *omitted variable bias* (Kennedy, 2008): country income per person (gross domestic product per capita, Chiu, 2010), family income inequality in a country (Gini index, Chiu, 2015), family SES (Chiu & Chow, 2010), home educational resources (Chiu & McBride-Chang, 2010), student gender (Chiu & McBride-Chang, 2006), reading self-concept (Chiu & Klassen, 2009), attitude toward reading (Chiu & Chow, 2015), parent attitude toward reading (Chiu & Chow, 2015), school climate (Thapa, Cohen, Guffey, & Higgins-D'Alessandro, 2013), and teacher gender (Winters, Haight, Swaim, & Pickering, 2013).

Methods

Like earlier studies on international data with similar structures and formats (e.g., Chiu & Chow, 2015), this study uses comparable methods. However, this study tests a different set of hypotheses, uses a later data set, and has different explanatory variables (e.g., variance of classmate family SES).

Data

In 40 countries in 2006, the International Association for the Evaluation of Education Achievement Progress in International Reading Literacy Study (IEA-PIRLS) assessed 208,057 fourth-grade students, and they, their parents, their teachers, and their school principals completed questionnaires. International experts from these countries defined reading achievement, built assessment frameworks, created test items, forward- and backward-translated them, and pilot-tested them to check their validity and reliability (for sample items and other details, see Martin, Mullis, & Kennedy, 2007, and www.pirls.org). Participating students completed an 80-min assessment booklet and then a 15–30-min questionnaire. Participating countries included Austria, Belgium-Flemish, Belgium-French, Bulgaria, Canada, Chinese Taipei, Denmark, England, France, Georgia, Germany, Hong Kong, Hungary, Iceland, Indonesia, Iran, Israel, Italy, Kuwait, Latvia, Lithuania, Luxembourg, Macedonia, Moldova, Morocco, Netherlands, New Zealand, Norway, Poland, Qatar, Romania, Russian Federation, Scotland, Singapore, Slovak Republic, Slovenia, South Africa, Spain, Sweden, Trinidad and Tobago, and the United States. The United States had country-wide missing data, so it was removed from our analysis. The World Bank (2007) collected economic data for each country (annual income and income inequality).

Methodological Design

Testing the above hypotheses across many countries and schools requires representative samples, precise achievement tests and questionnaires, and appropriate statistical models. IEA chose at least 150 representative schools in each country, based on neighborhood SES and student intake. From each school, IEA selected one or two 4th grade classes, yielding a sample size of at least 4,000 students per country (stratified sampling, Martin et al., 2007). Then, IEA applied suitable weights to create representative samples of each country's schools and 4th grade students. Students who refused to take the exam, could not physically take it, had intellectual disabilities, or did not understand the test language constituted altogether less than 4% of the original sample.

IEA gave the students reading *subtests* (overlapping subsets from the total set of all open-ended and multiple choice questions, namely a *balanced incomplete block* test; F. B. Baker & Kim, 2004). These subtests enable wider coverage of reading skills, reduce student fatigue, and reduce test-learning effects (F. B. Baker & Kim, 2004). To estimate each student's reading competence with greater precision, a graded-response, Rasch model measured the difficulty of each test item (F. B. Baker & Kim, 2004).

We reduced measurement error by using multiple questionnaire items for each theoretical construct (e.g., SES) and creating corresponding Rasch model-based indices. In each country, multigroup Rasch models for each item yielded similar parameters, which indicates measurement equivalence across countries (May, 2006). (In contrast to factor analysis, a multi-group Rasch model has the advantages of requiring only one invariant anchor item across countries and modeling heterogeneous use of the ordinal rating scale; Rossi, Gilula, & Allenby, 2001). All reliabilities were measured with the composite score reliability coefficient, which is more precise than Cronbach's alpha (Rowe & Rowe, 1997). Other studies have also shown consistent questionnaire responses and participant understandings across countries (Brown, Micklewright, Schnepf, & Waldmann, 2007; Martin, Mullis, & Kennedy, 2003, 2007; Schulz, 2003).

Multilevel analysis of plausible values yields standard errors that are more precise than those of ordinary least squares (Goldstein, 2011; Wu, 2005). Also, missing questionnaire response data

(8%) can bias results, reduce estimation efficiency, and complicate data analyses. To estimate these missing data, we use Markov Chain Monte Carlo multiple imputation, which is more effective than deletion, mean substitution, or simple imputation (Peugh & Enders, 2004).

Variables

Reading achievement is the outcome variable. Explanatory variables are added at the country, classroom, and student levels. See Table 1 for summary statistics.

Reading achievement. The PIRLS framework defines reading literacy as the ability to understand and use written language forms required by society and/or valued by individuals (Mullis, Kennedy, Martin, & Sainsbury, 2006). Young children often read for two major purposes: (a) for literary experience and (b) to acquire and use information. Hence, the test questions are divided equally to address each purpose (50% each). PIRLS considers four comprehension processes: attend to and retrieve explicitly stated information (20% of the questions); make simple inferences (30%); interpret and integrate information and ideas (30%); and evaluate content, language, and textual elements (20%). There were 64 multiple-choice items and 62 constructed-response items (interrater agreement on participant responses to the latter was 90%; Martin et al., 2007).

Parent rating of student's past literacy skills. To reduce measurement error, a graded response Rasch-based index was created from a parent's or guardian's responses to multiple survey items regarding a student's literacy skills before entering elementary school (using the Warm [1989] procedure). These items were as follows: my child (a) recognizes most of the alphabet letters, (b) reads words, (c) reads sentences, (d) writes letters of the alphabet, and (e) writes words. The possible choices "were not at all," "not very well," "moderately well," and "very well." Reliability was very high: 0.95.

Country. Country variables were (a) economic growth, (b) family income inequality and (c) distribution of students across schools. The World Bank (2007) measured economic growth via gross domestic product per capita (*GDP per capita*) and family income inequality via GDP Gini. *GDP Gini* is the difference of the integral of the cumulative distribution function of a perfectly equal income society minus the integral of the cumulative distribution function of the actual society's income. GDP Gini can range from 0 (perfect equality; everyone has equal income) to 100 (perfect inequality; one person has all the income, and everyone else's income is zero). For non-normal distributions (e.g., household income), the Gini index is especially suitable (McKenzie, 2005).

School clustering indicates the degree of school streaming within a country. We compute it by the ratio of variance of students' *past literacy skills* (reported by parents before schooling) across schools divided by the country variance of students' *past literacy skills*.

Family. Family variables include family SES, home educational resources, and parent attitude toward reading. The *family SES* index was created from father's education, mother's education, father's occupation, mother's occupation, and family financial situation. Occupation responses were recoded according to job status (according to Ganzeboom, De Graaf, & Treiman, 1992). Family financial situation was measured with the question "How well off do you think your family is financially?" and the choices *not at all well-off, not very well-off, average, somewhat well-off,* and *very well-off.* The reliability of the family SES index was 0.94.

The *home educational resources* index was created from responses to the following questions about availability at home of the following educational resources: (a) computer, (b) study desk/table for own use, (c) students' own books, (d) daily newspaper access, (e) number of books at home; and (f) number of children's books at home. For (a)–(d), the choices were yes or no. For (e)-(f), the choices were 0 to 10; 11 to 25; 26 to 100; 101 to 200; and more than 200. The reliability of this index was 0.79.

The *parent attitude toward reading* index was created from the following items: (a) I read only if I have to; (b) I read only if I need information; (c) I like talking about books with other people; (d) I like to spend my spare time reading; and (e) reading is an important activity in my home. The choices were *disagree a lot, disagree a little, agree a little,* and *agree a lot.* The reliability of this index was 0.82.

Classmates' families. To test whether classmates' characteristics were linked to a student's reading achievement, we included both a student characteristic (e.g., *SES*) and its class mean (e.g., *class mean SES*) in a regression. Controlling for a student's own *family SES*, the regression coefficient of *class mean SES* indicated the link between classmates' SES and a student's reading achievement. Similarly, *class mean parent attitude toward reading* and *class mean home educational resources* were entered with *parent attitude toward reading* and *home educational resources*, respectively.

School and teacher. School and teacher variables were school violence, female teacher, and homework mismatch. *School violence* was an index created from the items (a) I was bullied by another student; (b) someone in my class was bullied by another student; (c) I was injured by another student; (d) someone in my class was injured by another student during the last month at school. Choices were yes or no. The reliability of this index was 0.93.

Female teacher indicated whether the teacher was female (or male). *Homework mismatch* serves as an inverse proxy for teacher understanding of students, and is computed for each teacher as follows: [the sum of (*Teacher's expected daily reading time for homework by his or her students* – a *student's reported daily reading time for homework*)^2 for each student of this teacher] divided by the total number of these students. *Teacher's expected daily reading time for homework by his or her students* and *student's reported daily reading time for homework* were both on a 4-point scale with the following choices: (a) never have reading HW, (b) half hour or less, (c) half hour and one hour, and (d) one hour or more.

Student. Student variables were girl, reading attitude, and reading self-concept. *Girl* indicated whether the student was female (or male). *Student reading attitude* was an index created from the following items: (a) I read only if I have to, (b) I like talking about books with other people, (c) I would be happy if someone gave me a book as a present, (d) I think reading is boring, and (e) I enjoy reading. Choices were *disagree a lot, disagree a little, agree a little* and *agree a lot.* The reliability of this index was 0.86.

Student reading self-concept was an index created from the following items: (a) reading is very easy for me, (b) I do not read as well as other students in my class, and (c) reading aloud is very

Table 1
Summary Statistics of Variables (N = 208,044)

Variable	Mean	*SD*	Description
Reading achievement	496	113	The student reading scores estimated by the Rasch models were calibrated to a mean of 500 and a standard deviation of 100 (Martin, Mullis, & Kennedy, 2007). Min = 5, Max = 813.
Past literacy skills variable			
Students' past literacy skills	.00	1.00	Index of students' ability at 1st grade as rated by parents: Recognize most of the letters of the alphabet; Read some words; Read sentences; Write letters of the alphabet; Write some words. Choices = not at all, not very well, moderately well, very well. Reliability = .95. Min = −2.49, Max = 1.82.
Country level variables (**Country**)[a]			
Log GDP per capita	9.73	.65	Min = 7.97, Max = 10.50. World Bank, 2007. Linear GDP per capita fit the data worse.
GINI	35.44	7.90	Scores range from 0 (perfect equality; same incomes for all) to 100 (perfect inequality; one person has all the income). Min = 25, Max = 58. World Bank, 2007.
Clustering of students across schools by past literacy skills reported by parents	.12	.09	Ratio of variance of *past literacy skills reported by parent* across schools divided by the country variance of *past literacy skills reported by parent.* Min = .05, Max = .54.
Family variables at the student level (**Family**)			
Family SES	.00	1.00	Index of: Father education; Mother's education; Father's occupation; Mother's occupation; Family's financial situation. Reliability = .94. Min = −2.95, Max = 2.84.
Parents' attitude towards reading	.00	1.00	Index of: I read only if I have to; I like talking about books with other people; I like to spend my spare time reading; I read only if I need information; Reading is an important activity in my home. Choices: *disagree a lot, disagree a little, agree a little, agree a lot.* Reliability = .82. Min = −2.94, Max = 1.69.
Home educational resources	.01	1.00	Index of: Number of books at home; Number of children books at home; Have a computer at home; Have a student desk/table for my use at home; Have books of my very own at home. Choices: 0–10, 11–25, 26–100, 101–200, >200 for number of books; 0–10, 11–25, 26–50, 51–100, >100 for number of children books; yes, no for others. Reliability = .76. Min = −2.54, Max = 1.98.
Class mean family variables at the class level (**Classmates_Families**)			
Class mean SES	.00	.61	Classmates' family SES. Min = −2.72, Max = 2.18.
Class mean home educational resources	.01	.68	Classmates' home educational resources. Min = −2.54, Max = 1.98.
Class mean parents' attitude towards reading	.00	.37	Class mean of parents' attitude towards reading. Min = −2.67, Max = 1.41.
School variables at the student level (**School**)			
School violence	.00	1.00	Index of: I was bullied by another student; Someone in my class was bullied by another student; I was injured by another student; Someone in my class was injured by another student during the last month at school. Choices: yes, no. Reliability = .93. Min = −1.37, Max = 2.09.
Teacher variables at the teacher level (**Teacher**)			
Female teacher	.84		1 = female teacher; 0 = male teacher
HW mismatch	.89	.61	A measure of teacher misjudgment of students. For each teacher, compute Σ (Teacher reported daily reading HW time − Student reported daily reading HW time)^2 / total # students for this teacher, where Teacher reported HW time and student reported HW time are in 4-point scale. Choices: *never have reading HW, half hour or less, half hour and one hour, one hour or more.* Min = 0, Max = 8.19.
Student variable at the student level (**Student**)			
Girl	.49		1 = Girl; 0 = Boy.
Students' reading attitudes	.00	1.00	Index of: I read only if I have to; I like talking about books with other people; I would be happy if someone gave me a book as a present; I think reading is boring; I enjoy reading. Choices = *disagree a lot, disagree a little, agree a little, agree a lot.* Reliability = .79. Min = −3.00, Max = 1.59.
Students' reading self-concept	.00	1.00	Index of: Reading is very easy for me; I do not read as well as other students in my class; When I am reading by myself, I understand almost everything I read; I read slower than other students in my class. Choices = *disagree a lot, disagree a little, agree a little, agree a lot.* Reliability = .80. Min = −2.77, Max = 1.54.
Class mean of student variables at the class level (**Classmate**)			

(*table continues*)

Table 1 (*continued*)

Variable	Mean	*SD*	Description
Class mean past literacy skills	.00	.48	Classmates' past literacy skills. Min = −2.49, Max = 1.66.
Class mean attitude towards reading	.00	.41	Class mean of students' attitude towards reading. Min = −1.66, Max = 1.59.
Class variance variable at the class level (**Classmates_Variance**)			
Past literacy skills—class variance	.80	.35	Variance of classmate past literacy skills. Min = 0, Max = 3.89.
SES -class variance	.66	.31	Variance of classmate family SES. Min = 0, Max = 3.35.
Students' attitude towards reading—class variance	.87	.35	Variance of classmate attitude towards reading. Min = 0, Max = 3.54.

Note. Indices were standardized (M = 0; SD = 1). All reliabilities refer to the composite score reliability coefficient.
[a] Bold letters indicate vectors (e.g., **Country**) in their order of entry into the regression.

hard for me. The choices were *disagree a lot*, *disagree a little*, *agree a little*, and *agree a lot*. As the reliability of this index was 0.62, results involving this index require cautious interpretation.

Classmates. As with classmate family variables, we model classmate attributes by including both a student characteristic (e.g., *attitude toward reading*) and its class mean (e.g., *class mean attitude toward reading*) in a regression. We apply this procedure for *class mean attitude toward reading* and *class mean parent rating of past literacy skills*.

Variance among classmates. To test whether the distribution of students is related to student reading achievement, we create classroom variance and standard deviation (*SD*) variables (we retain the explanatory variable [variance vs. *SD*] that accounts for more outcome variance). For each class, we compute and store the school variance of family SES in the variable *variance of classmate SES*. Likewise, we create the variables *variance of classmate past literacy skills* (as perceived by their parents) and *variance of classmate attitude toward reading*. We also compute the square root of each variable to obtain *SD* of classmate SES, *SD of classmate past literacy skills*, and *SD of classmate attitude toward reading*.

Analysis

A variance components model tested for significant differences at each level (Goldstein, 2011).

$$\mathbf{Reading}_{ijk} = \beta + e_{ijk} + f_{jk} + g_k \quad (1)$$

The outcome variable $\mathbf{Reading}_{ijk}$ of student i in school j in country k has a grand mean intercept β, with unexplained components (*residuals*) at the student, school, and country levels (e_{ijk}, f_{jk}, g_k).

Explanatory variables were entered in sequential sets to estimate the variance explained by each set (Kennedy, 2008). As student current reading achievement is often related to past reading achievement, we control for this variable to narrow the analysis to a shorter time frame between past and current reading achievement. Hence, we first enter a student's past literacy skills before enrolling in school, as reported by a parent.

Next, country variables might affect family variables; for example, in a wealthy country, a family often has more opportunities to earn more money. As families might choose their children's schools, family and classmate family variables might be related to school variables. Also, attributes of school variables might influence teacher behaviors. For example, school violence might influence a female teacher's relationships with her students. All of these might affect students. After controlling for all of these variables, we test our hypotheses involving classmates and classmate variance. While this conservative approach might yield false negatives, we have greater confidence in significant links between classmate characteristics and student reading achievement. Hence, we entered the variables as follows: past literacy skills, country, family, school, teacher, student, classmates, and classmate variance (see variable descriptions in Table 1). All continuous variables were centered on their country mean.

For some explanatory variables, small values might have positive links to reading achievement while large values have negative links to it (or vice versa), so we test for nonlinear relations by adding the quadratic or log term of each explanatory variable. For example, in addition to *school clustering*, we also test *school clustering*^{2} or *log (school clustering)*. If the log term accounts for more variance than the linear term, it suggests *diminishing marginal returns* (e.g., a thirsty person benefits more from the first glass of water than from the last glass of water; an additional book benefits a poorer student with few books more than a richer student with many books; Chiu, 2015).

We began by entering past student achievement (*Past_Literacy_Skills*).

$$\begin{aligned}\mathbf{Reading}_{ijk} = {} & \beta_{000} + e_{ijk} + f_{0jk} + g_{00k} + \beta_{rjk}\mathit{Past\_Literacy\_Skills}_{ijk} \\ & + \beta_{00s}\mathbf{Country}_{00k} + \beta_{tjk}\mathbf{Family}_{ijk} \\ & + \beta_{ujk}\mathbf{Classmates\_Families}_{ijk} \\ & + \beta_{vjk}\mathbf{School\ \&\ Teacher}_{ijk} + \beta_{wjk}\mathbf{Student}_{ijk} \\ & + \beta_{xjk}\mathbf{Classmates}_{ijk} + \beta_{zjk}\mathbf{Classmates\_Variance}_{ijk}\end{aligned} \quad (2)$$

Then, we entered a vector of s variables at the country level: GDP per capita, GDP Gini, school clustering of students by past literacy skills (**Country**, see Table 1). We tested whether sets of predictors were significant with a nested hypothesis test (χ^2 log likelihood, Kennedy, 2008). Nonsignificant variables were removed. This specification tests whether more school clustering is linked to higher reading achievement (H-1a *customized instruction*). To test for nonlinear relations (H-1b *help opportunities*), we use quadratic terms (e.g., *Gini*^{2}) or logarithmic terms (*log* [GDP per capita]).

Next, we applied the procedure for **Country** to family variables: family SES, parents' attitude toward reading, and home educational

resources (**Family**). The additional variables of home educational resources and parent attitude toward reading help distinguish between tangible and intangible resources to test whether they have different links to reading achievement (Chiu, 2007).

We also applied the procedure for **Country** to the class means of family SES, home educational resources, and parents' attitudes toward reading (**Classmates_Families**). This specification tests whether higher classmate SES is linked to higher reading achievement (H-7 *share*).

Then, we applied the procedure for **Country** to school and teacher variables: school violence, female teacher, and homework mismatch (**School & Teacher**). Next, we applied the above procedures for **School & Teacher** on the student variables: gender, student's attitude toward reading, student's reading self-concept (**Student**).

Then, we applied the above procedures for **Student** to the class mean of past literacy skills and class mean of students' attitude toward reading (**Classmates**). This specification tests the competing hypotheses of whether a student's higher reading achievement is linked to classmates' *lower* past reading achievement (H-5a *social comparison*) or *higher* past reading achievement (H-5b *share*). It also tests whether classmates with superior reading attitudes are linked to higher reading achievement by a student (H-6 *modeling, norms*).

Lastly, we applied the above procedures for **Classmates** on the class variances of past reading achievement, SES, attitude toward reading (**Classmates_Variance**)m, along with their quadratic terms to test for nonlinear relations. This specification tests whether higher academic achievement is linked to higher variance in classmate past achievement (H-2a *help opportunities*), to extremely high variance in it (H-2b *homophily*), to higher variance in classmate SES (H-3 *diversity*), to higher variance in classmate reading attitude (H-4a *contrasting cases*), or to extremely high variance in it (H-4b *homophily*). In an alternate specification, we also tested classmate standard deviations and kept the significant variables (variance or *SD*) that accounted for more of the differences in reading achievement.

To determine whether high-achieving versus low-achieving students have greater access to or benefits from available resources, we tested if the above links differed across student subsamples. We created six subsamples of students by reading achievement (bottom 10%, bottom 20%, bottom 50%, top 50%, top 20%, and top 10% from each country). These specifications test hypotheses regarding differences in relations across students by past achievement (H-2b *customized instruction*; H-3 *Matthew effect*; H-5c *help opportunities*; H-6 *modeling, norms*).

We reported how a 10% increase in each continuous variable above its mean was linked to reading achievement (result = b * *SD* * [10%/34%]; 1 *SD* $\approx$ 34%). The percent increase is not linearly related to standard deviation, so scaling is not warranted.

We used an alpha level of .05. To control for the false discovery rate, we used the two-stage linear step-up procedure, which outperformed 13 other methods in computer simulations (Benjamini, Krieger, & Yekutieli, 2006). Using standardized scores within each country, we repeated the analyses. The small sample of countries (N = 40) limits identification of nonsignificant country-level results (for a 0.4 effect size at p = .05, statistical power = 0.74; Konstantopoulos, 2008; see Appendix A for details). We tested whether the results differed across the 40 countries via multilevel analysis. Also, we analyzed residuals for influential outliers. The above analyses were completed with ICL (Hanson, 2002), MLwin (Rasbash, Steele, Browne & Goldstein, 2015) and LISREL (Jöreskog & Sörbom, 2012).

Results

Summary Statistics

This sample included a variety of countries. They ranged from poor, very unequal nations (e.g., Iran) to rich, relatively equal ones (e.g., Luxembourg). See Table 1 for overall summary statistics (see Appendix B for correlation-variance-covariance matrices).

Explanatory Model

Past literacy skills, country, family, school, teacher, student, classmates and class variance variables accounted for differences in students' reading achievement (see Tables 2 and 3 and Appendix C for graphs of nonlinear relations). The variances in reading achievement at the country, school, and student levels were 58%, 13%, and 29%. All results discussed below describe first entry into the regression, controlling for all previously included variables. Ancillary regressions and statistical tests are available upon request.

Past literacy skills. Students whose parent-reported past literacy skills before schooling exceeded the mean by 10% averaged 5 more points in reading achievement (Table 2, Model 1). Past literacy skills accounted for 2% of the differences in students' reading achievement.

Country. Students in richer or more equal countries scored higher (Table 2, Model 2). If a richer country's GDP per capita exceeded the country mean by 10%, its students averaged 4 more points in reading achievement. (Regressions with linear GDP per capita did not fit the data as well, explaining less of the variance in students' reading achievement.) When a country's GINI exceeded the mean by 10%, its students averaged 20 points lower in reading.

If a country's clustering of students based on past literacy skills exceeded the mean by 10%, students averaged 45 more points in reading achievement, supporting hypothesis H-1a (*customized instruction*). Furthermore, school clustering showed a nonlinear relation with reading achievement (for graphs of all nonlinear relations, see Appendix C). This relation was highest when school clustering was 0.30 (~95th percentile of a normal curve or 2.0 SDs above the mean; 2.0 = [0.30 − 0.12]/0.09 = ([X−mean]/*SD*). Extreme school clustering beyond two standard deviations (>0.30) was associated with lower student reading achievement, supporting hypothesis H-1b (fewer *help opportunities*) and rejecting *homophily* as a possible explanation (which would have predicted higher reading achievement). Together, country variables accounted for about 29% the total variance in reading achievement and for 51% of its differences across countries.

Family. Family variables were linked to students' reading achievement (Table 2, Model 3). When students had 10% more home educational resources than average, they scored 4 points higher in reading, on average. When parents had 10% better reading attitude than average, their children averaged 2 points

Table 2
Summaries of 9 Multilevel Regressions Predicting Students' Reading Achievement, With Unstandardized Coefficients (and Standard Errors in Parentheses)

	Regressions Predicting Reading Achievement								
Explanatory variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Past literacy skills	15.370*** (.136)	15.448*** (.130)	14.308*** (.129)	13.397*** (.129)	13.252*** (.137)	13.154*** (.136)	9.155*** (.113)	9.647*** (.125)	9.053*** (.118)
Log GDP per capita		39.373*** (3.911)	33.660*** (3.610)	21.877*** (3.508)	23.029*** (3.101)	21.418*** (3.212)	21.138*** (3.454)	22.498*** (3.408)	25.029*** (3.342)
GINI		−8.532*** (.170)	−8.455*** (.158)	−7.221*** (.141)	−7.066*** (.144)	−7.632*** (.145)	−7.602*** (.137)	−7.596*** (.135)	−7.032*** (.129)
Clustering of students by past literacy skills^2		−2749.936*** (148.694)	−2573.550*** (135.402)	−2276.796*** (124.325)	−2138.506*** (117.597)	−2217.040*** (121.620)	−2060.949*** (115.170)	−2052.441*** (119.183)	−1913.587*** (118.780)
Clustering of students by past literacy skills		1718.178*** (75.372)	1676.304*** (70.676)	1294.815*** (65.601)	1347.410*** (58.937)	1356.799*** (59.620)	1311.454*** (64.887)	1150.817*** (62.395)	1100.190*** (62.993)
Family SES			15.429*** (.143)	10.379*** (.153)	9.706*** (.152)	9.748*** (.144)	8.138*** (.141)	8.017*** (.153)	8.072*** (.147)
Home education resources			15.153*** (.152)	12.725*** (.152)	13.780*** (.149)	12.790*** (.157)	8.587*** (.144)	8.621*** (.146)	8.870*** (.149)
Parents' attitude towards reading			7.251*** (.126)	7.075*** (.120)	6.777*** (.130)	6.723*** (.124)	4.977*** (.119)	4.941*** (.108)	5.293*** (.116)
Class mean SES^2				−6.226*** (.644)	−6.483*** (.634)	−6.413*** (.678)	−5.137*** (.609)	−4.924*** (.607)	−5.473*** (.613)
Class mean SES				24.751*** (.818)	23.228*** (.874)	24.359*** (.855)	22.600*** (.746)	21.915*** (.784)	20.952*** (.728)
Class mean home education resources^2				2.704*** (.635)	2.635*** (.609)	2.753*** (.614)	2.029*** (.603)	1.422* (.625)	1.579** (.565)
Class mean home education resources				23.694*** (.868)	23.105*** (.904)	22.534*** (.849)	23.110*** (.822)	23.034*** (.895)	22.372*** (.831)
Class mean past literacy skills^2				18.651*** (.737)	18.283*** (.720)	16.996*** (.726)	15.371*** (.697)	14.486*** (.758)	14.966*** (.672)
Class mean past literacy skills				10.529*** (.789)	10.220*** (.794)	11.232*** (.717)	11.106*** (.707)	10.773*** (.755)	10.363*** (.726)
School violence					−6.129*** (.125)	−5.573*** (.119)	−3.549*** (.116)	−3.221*** (.120)	−3.43086*** (.110)
Female teacher					4.629*** (.623)	4.428*** (.612)	4.387*** (.596)	4.484*** (.552)	4.017*** (.609)
HW mismatch					−5.373*** (.416)	−5.255*** (.379)	−4.826*** (.397)	−4.896*** (.400)	−5.293*** (.365)
Girl						14.220*** (.250)	9.145*** (.223)	9.642*** (.237)	9.596*** (.230)
Students' attitude towards reading							8.490*** (.119)	8.304*** (.115)	8.248*** (.119)
Students' reading self-concept reading							18.761*** (.117)	21.396*** (.111)	21.574*** (.107)
Class mean students' reading attitude^2								−1.922* (.866)	.690 (.831)
Class mean students' reading attitude								6.330*** (.679)	9.880*** (.691)
SES—class variance									2.503** (.765)

(*table continues*)

Note. GDP = gross domestic product; SES = socioeconomic status. Each regression included a constant term.
$^* p < .05$. $^{**} p < .01$. $^{***} p < .001$.

Table 2 (*continued*)

Explanatory variable	Regressions Predicting Reading Achievement								
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Past literacy skills—class variance^2									-5.639^{***} (.935)
Past literacy skills—class variance									13.461^{***} (2.088)
Students' reading attitude—class variance^2									-7.997^{***} (1.128)
Students' reading attitude—class variance^2									22.466^{***} (2.371)
	Explained variance at each level								
Variance at each level									
Country (58%)	.000	.506	.559	.616	.628	.628	.633	.635	.639
School (13%)	.031	.031	.189	.371	.387	.393	.418	.421	.429
Student (29%)	.048	.048	.186	.191	.196	.219	.289	.289	.289
Total variance explained	.018	.313	.403	.461	.471	.479	.505	.507	.510

Note. GDP = gross domestic product; SES = socioeconomic status. Each regression included a constant term.
$^* p < .05$. $^{**} p < .01$. $^{***} p < .001$.

higher in reading. Family variables accounted for an extra 7% of the variance in students' reading achievement.

Classmates' families. Classmate family variables were also linked to students' reading achievement (Table 2, Model 4). If classmate family SES exceeded the mean by 10%, students averaged 4 points higher in reading achievement, and this nonlinear relation was smaller for higher classmate family SES (*diminishing marginal returns*, Chiu, 2015; see Appendix C), supporting the classmate sharing hypothesis H-7. Moreover, students with classmates who had 10% more education resources at home averaged 5 points higher reading achievement, and this technically nonlinear relation was nearly linear (see Appendix C).

School and teacher. School and teacher variables were linked to students' reading achievement (Table 2, Model 5). Students in schools with 10% less school violence averaged 2 points higher in reading. Students with a female teacher averaged 5 points higher than those with a male teacher. Also, if teachers' misjudgment of students' homework time exceeded the mean by 10%, their students averaged 1 point lower in reading. School and teacher variables accounted for an extra 1% of the variance in students' reading achievement.

Student. Girls outscored boys by 13 points on average (Table 2, Model 6). Moreover, students with 10% better reading attitude or 10% higher reading self-concept averaged 3 or 6 points higher in reading, respectively (Table 2, Model 7). Student variables accounted for an extra 4% of the variance in students' reading achievement.

Classmates. When classmates' past literacy skills (parent-reported) exceeded the mean by 10%, students averaged 2 points higher reading achievement, supporting hypothesis H-5b (*share*) and rejecting hypothesis H-5a (*social comparison*). This nonlinear relation has a minimum when classmates' past literacy skills are -0.30 ($\sim$10th percentile at -0.63 *SD* [below the mean]; $-0.63 = [-0.30 - 0]/0.48$); at lower values beyond this turning point, student reading achievement scores were often higher (see Appendix C). This result shows that having more classmates whose past literacy skills differ from that of a student is linked to the latter's greater reading achievement, supporting hypothesis H-5C (*help opportunities*). Together, classmates' family variables accounted for an extra 6% of the variance in students' reading achievement.

When classmates' reading attitudes exceeded the mean by 10%, students averaged 1 point higher in reading (Table 2, Model 8), supporting hypothesis H-6 (*modeling, norms*); the link is slightly stronger for higher classmate reading attitudes (see Appendix C). This classmate variable accounted for an extra 0.1% of the variance in students' reading achievement.

Classmate variance. Classmate variances were also linked to student reading achievement, accounting for more of it than classmate standard deviations (Table 2, Model 9). When variance of classmates' past literacy skills exceeded the mean by 10%, students averaged 1 point higher in reading achievement, supporting hypothesis H-2a (*help opportunities*) and rejecting a *homophily* view. The maximum size of this link was at a variance of 1.19 ($\sim$86th percentile for a normal curve or 1.11 standard deviations above the mean; $1.11 = [1.19 - 0.80]/.35 = ([X-\text{mean}]/SD)$. Beyond this turning point, extreme variance was linked to lower reading achievement, supporting hypothesis H-2b (less *customized*

Table 3
Graphic Summary of Complex Regression Results

Explanatory variable	Overall result	Similar results for achievement subsamples?
School clustering by past literacy		Yes
Classmates parents'		
SES		Yes
SES variance		Only significant for top 50% and 20%
Home educational resources		Yes
Classmates'		
Past achievement		Only for bottom 10%, 20%, 50%, and top 50%; see text for top 20% and 10%
Past achievement variance		Only for bottom 10%, 20% and 50%; see text for top 50%, 20% and 10%
Reading attitude		Only significant for bottom 10%, 20% and 50%
Reading attitude class variance		Yes

Note. SES = socioeconomic status.

instruction; for higher achieving students, see the section Differences Across Achievement Subsamples below, and Appendix C).

Also, students in classes with 10% more variance in classmates' reading attitudes averaged 2 points higher in reading, supporting hypothesis H-4a (*contrasting cases*). The maximum size of this link was at a variance of 1.40 (~94th percentile for a normal curve or 1.51 standard deviations above the mean; 1.51 = [1.40 − 0.87]/.35 = ([X−mean]/*SD*)) and then is lower at higher variances, supporting hypothesis H-4b (*homophily*). Lastly, students in classes with 10% more SES variance averaged 0.2 points higher in reading achievement, supporting H-3 (*diversity*). Class variances accounted for an extra 1% of the variance in students' reading achievement. Other variables were not significant.

Differences Across Achievement Subsamples

Many variables showed similar, significant results in all achievement subsamples (see Table 4). These variables included past literacy skills, log GDP per capita, GINI, clustering of students by past literacy skills, parents' attitude toward reading, home educational resources, classmates' mean SES, classmates' home educational resources, school violence, students' gender, students' reading self-concept, and class variance in students' attitude toward reading. Specifically, the consistent results of school streaming on reading achievement across all achievement subsamples reject the views of labeling or social comparison, which would predict opposite relations for high- versus low-achieving students. (The consistency of nonlinear results can be seen by graphing the quadratic functions within the minimum and maximum values. Due to space considerations, these graphs are not included but are available upon request.)

The achievement subsample results varied for the following attributes: classmate reading attitude, classmate SES variance, mean and variance of classmates' past literacy skills, teacher gender, and homework mismatch. The classmate reading attitudes results were significant for only the lowest-achieving 10%, 20% and 50% of students, supporting hypothesis H-6 (*modeling, norms*). Furthermore, variance in classmates' SES was significant for only the top 50% and top 20% of students by achievement, suggesting that students with higher past literacy skills benefit more than other students from classmates' diverse resources and experiences; the results are not significant for the other subsamples. These results generally support hypothesis H-3 (*Matthew effect*) except for the surprising nonsignificant result for the top 10% of students.

For most subsamples of students, student reading achievement's links to the mean and variance of classmates' past literacy skills were similar to the overall result. However, the highest-achieving 20% and 10% of students had higher reading achievement when their classmates had *lower* past literacy skills, consistent with hypothesis H-5c (*help opportunities*). Also, the variance of classmates' past literacy skills was not significantly related to reading achievement for the highest-achieving 50% of students, and negatively related for the highest-achieving 20% and 10% of students, supporting H-2b (less *customized instruction*).

Teacher gender and homework mismatch were not significant for the top 10% of students in each country, suggesting that these factors do not affect the highest-achieving students. Perhaps higher

Table 4

Multilevel Regressions of Six Sub-Samples by Achievement, the Bottom 10%, 20%, and 50%, and the Top 10%, 20%, and 50%

	Regressions of achievement subsamples					
	Bottom			Top		
Explanatory variable	10%	20%	50%	50%	20%	10%
Past literacy skills	2.700*** (.267)	3.178*** (.197)	4.724*** (.151)	3.832*** (.111)	2.219*** (.134)	1.933*** (.162)
Log GDP per capita	19.885*** (2.312)	17.434*** (2.127)	12.169*** (2.244)	41.222*** (1.948)	40.106*** (1.825)	34.197*** (2.124)
GINI	−2.526*** (.170)	−3.524*** (.171)	−3.614*** (.168)	−6.288*** (.078)	−4.940*** (.076)	−3.149*** (.095)
Clustering of students by past literacy skills²	−1443.740*** (90.013)	−1493.313*** (83.816)	−1419.103*** (72.642)	−1878.351*** (72.688)	−1620.417*** (78.178)	−1425.541*** (75.392)
Clustering of student by past literacy skills	480.530*** (46.230)	524.279*** (43.564)	574.523*** (39.185)	1122.714*** (37.471)	900.990*** (32.955)	786.002*** (40.344)
Family SES	2.201*** (.311)	2.459*** (.221)	3.636*** (.164)	4.263*** (.127)	2.541*** (.160)	1.768*** (.217)
Home education resources	2.114*** (.279)	3.208*** (.238)	7.079*** (.169)	4.092*** (.144)	3.166*** (.182)	3.082*** (.241)
Parents' attitude towards reading	.609** (.212)	.999*** (.174)	1.729*** (.138)	1.674*** (.114)	.428 (.150)	.448* (.182)
Class mean SES²	.365 (.535)	.207 (.487)	−1.748*** (.517)	.869* (.441)	3.134*** (.519)	2.450*** (.524)
Class mean SES	7.749*** (.773)	7.688*** (.696)	6.981*** (.582)	10.834*** (.568)	6.714*** (.611)	5.180** (.697)
Class mean home education resources²	−.064 (.586)	.558 (.532)	1.210* (.490)	−1.740*** (.475)	−3.713*** (.486)	−3.585*** (.544)
Class mean home education resources	8.573*** (.931)	11.145*** (.828)	14.213*** (.710)	10.067*** (.624)	8.348*** (.698)	7.502*** (.798)
Class mean past literacy skills²	2.246** (.773)	3.171*** (.714)	4.588*** (.603)	5.203*** (.592)	−.869 (.541)	−1.737** (.613)
Class mean past literacy skills	−4.212*** (.948)	−1.789* (.796)	2.777*** (.608)	1.150* (.541)	−2.016*** (.551)	−.893 (.609)
School violence	−.545* (.220)	−.693*** (.160)	−1.902*** (.140)	−1.930*** (.113)	−1.150*** (.139)	−1.286*** (.184)
Female teacher	1.974** (.617)	1.895*** (.549)	3.578*** (.486)	2.131*** (.459)	1.709*** (.516)	−.711 (.550)
HW mismatch	−3.854*** (.449)	−4.590*** (.438)	−4.767*** (.368)	−1.533*** (.299)	−.821* (.330)	−.239 (.348)
Girl	1.148** (.420)	2.626*** (.332)	4.666*** (.246)	1.898*** (.196)	1.245*** (.285)	1.206*** (.314)
Students' attitude toward reading	.998*** (.226)	1.504*** (.189)	2.443*** (.135)	3.915*** (.117)	1.828*** (.148)	.925*** (.200)
Students' reading self-concept	2.790*** (.217)	5.348*** (.168)	9.244*** (.136)	9.220*** (.109)	5.825*** (.156)	4.150*** (.209)
Class mean students' attitude towards reading²	.292 (1.053)	.033 (.957)	1.492 (.784)	.052 (.725)	.571 (.799)	−1.490 (.797)
Class mean students' attitude towards reading	6.701*** (.844)	7.249*** (.688)	8.628*** (.625)	.223 (.540)	−.025 (.616)	−.124 (.665)
SES—class variance	.080 (.879)	.643 (.722)	.615 (.709)	1.443* (.591)	2.618*** (.641)	1.194 (.694)
Past literacy skills—class variance²	−7.573*** (.942)	−7.312*** (.830)	−9.954*** (.768)	1.514 (.805)	1.328 (.886)	1.373 (1.057)
Past literacy skills—class variance	16.555*** (2.012)	14.606*** (1.893)	20.035*** (1.741)	−2.653 (1.786)	−3.899* (1.914)	−5.607** (2.128)
Variance of classmates' attitude towards reading²	−9.913*** (1.045)	−8.755*** (.997)	−7.521*** (.972)	−2.641** (.967)	−2.174* (1.012)	−1.449 (1.210)
Variance of classmates' attitude towards reading	27.031*** (2.522)	21.891*** (2.227)	20.775*** (1.940)	8.061*** (1.948)	6.365** (2.237)	5.597* (2.574)
Variance at each level						
Country	.616	.628	.628	.633	.635	.639
School	.373	.389	.394	.418	.421	.429
Student	.284	.283	.294	.371	.372	.372
Total variance explained	.489	.498	.501	.529	.531	.534

Note. GDP = gross domestic product; SES = socioeconomic status. Each regression included a constant term.
* $p < .05$. ** $p < .01$. *** $p < .001$.

ability students can use their cognitive capabilities to learn regardless of the presence or absence of these resources (technically, their cognitive capabilities might *substitute* for learning resources, Mankiw, 2014).

None of the above results showed labeling or social comparison effects. Analyses using standardized scores within each country yielded similar results. Analyses of residuals showed no influential outliers.

Discussion

In this study, we tested whether greater homogeneities of students' past achievement (reported by parents), reading attitudes, or family SES was related to the reading achievement of 208,057 fourth grade primary school students in 40 countries. There are seven major results. First, greater homogeneity of students' parent-reported past literacy skills within a school was linked to higher student reading achievement (customized instruction), although this link was weaker for extremely homogeneous schools (fewer help opportunities). Second, greater differences in classmates' past literacy skills were linked to higher student reading achievement. However, extremely high differences weakened this link, especially for high-achieving students (fewer customized instruction opportunities). Third, when classmates included a mixture of different family SES, students had higher reading achievement (diversity), especially for high-achieving students (Matthew effect). Fourth, mixing students with different reading attitudes in a class was linked to higher reading achievement (contrasting cases), although extremely high variance of reading attitudes weakened this link (homophily). Fifth, a student with higher-achieving classmates often had higher reading achievement compared to other students (share), especially when classmates' past achievements differed substantially from that of the student (help opportunities). Sixth, low-achieving students whose classmates had better reading attitudes had higher reading achievement (modeling, norm). Lastly, students whose classmates had higher family SES had higher reading achievement (share), although with diminishing returns.

Past Achievement

The past achievement results suggest that greater homogeneity of students within a school (possibly due to streaming) facilitates customized instruction and that mixing students by past achievement in classrooms (possibly due to less tracking) increases help opportunities, both of which are linked to higher reading achievement. Specifically, greater school clustering of students by past literacy skills—up to about two standard deviations above the mean—might help educators target curricula and instruction to specific schools of students with similar academic competence to help them learn more (Watanabe, 2008). However, the lower reading achievement of students in education systems with extremely high clustering (more than two standard deviations above the mean) suggest that classmates with extremely similar past achievement reduce pairings of higher- and lower-achieving students, resulting in fewer help opportunities, fewer learning opportunities, and lower reading achievement (Chiu & Chow, 2015).

The results regarding classmate past achievement suggest that students benefited from the help opportunities and classmate sharing of ideas. When differences among classmates' past reading achievements were larger (less classroom tracking), students had higher reading achievement, suggesting that they had more help opportunities (as givers or recipients) and capitalized on them to learn more (Chiu & Chow, 2015). Likewise, when a student's classmates had very high or very low past literacy skills (not medium past literacy skills), he or she had higher reading achievement, consistent with the view that a student learns more by receiving help from higher-achieving classmates or by giving help to lower-achieving classmates (Guyon et al., 2012). For example, the highest-achieving 20% and 10% of students benefited primarily from opportunities to help others, so those who had more classmates with low past literacy skills might have had more helping opportunities—which would account for their higher reading achievement. This claim that a student's classmates use their knowledge to share ideas and provide help (rather than only socially compare, Liu et al., 2005) is further supported by the positive link between classmates' past literacy skills and student reading achievement.

When classmates' differences in past literacy skills were extremely high however, high-achieving students' reading achievement was lower, consistent with the view that their fewer customized instruction opportunities hindered their learning (Smith, 2013). As higher-achieving students acting as helpers resemble teachers, helping classmates with a wide range of past literacy skills might be too challenging (and potentially frustrating), in contrast to the easier task of helping classmates with a smaller range of past literacy skills (akin to customized instruction).

Together, these results suggest that moderate streaming across schools and moderate mixing of students by past achievement might enable both customized instruction and help opportunities. Clustering moderately similar students together within a school can encourage teachers to customize their instruction to students within a narrower range of competence. Likewise, a moderate mixture of classmates by past achievement can increase help opportunities via pairs of higher- and lower-achieving students. Placing high-achieving students among a limited range of low-achieving classmates can support the former's customized instruction. These results also suggest avoiding extremes, where severe school streaming reduces help opportunities and sharp differences among classmates prevent customized instruction by higher-achieving students.

Classmate Family SES and Home Educational Resources

In addition to classmates' past literacy skills, their higher family SES and home educational resources were also linked to higher student reading achievement. These results suggest that classmates shared educational materials, ideas, and experiences with a student to aid his or her reading achievement (e.g., Kerr, Pekkarinen, & Uusitalo, 2013). As classmate family SES and national income both show diminishing marginal returns on reading achievement, they benefit lower SES students more than higher SES students, consistent with earlier studies (e.g., Chiu, 2015). Furthermore, it suggests that mixing high SES students with low SES students in a classroom (diversity) would increase the benefits of classmate SES (Chiu, 2015), which is supported by our classmate family SES variance results; when

the family SES of classmates differed more, student reading achievement was higher. However, this greater diversity of classmate SES primarily benefits the highest-achieving 50% and 20%, with no significant relations for the other students (Matthew effect).

Together, these classmate family SES results suggest that mixing students of different SES in a classroom benefits lower SES students and the highest-achieving 50% and 20% of students. Moreover, the diminishing returns results show that lower SES students benefit more than higher SES students from the same resource, and suggest that mixing students by SES yields higher overall achievement, consistent with earlier studies (e.g., Chiu, 2010). If SES is highly correlated with student past achievement in a region or country, however, it may not be possible to both group students by similar past achievement into schools (streaming) and ensure that all students have some classmates with high SES or home educational resources. Also, future studies might consider which interventions might help lower-achieving students benefit from classmates with diverse family SES.

Classmate Reading Attitudes

Like the past literacy skills and family SES of classmates, their better reading attitudes and greater differences in reading attitude were both linked to higher student reading achievement. When classmates had better reading attitudes, students with lower past reading achievement (lowest 10%, 20%, and 50%) had higher reading achievement, suggesting a modeling/norm effect; higher achieving students did not benefit, possibly in part because they already had higher reading attitudes (reading attitude and reading achievement were weakly correlated, $r = .12$). Furthermore, when the reading attitudes of classmates differed more, students showed higher reading achievement, consistent with learning from contrasting cases (Schwartz & Bransford, 1998) of students with good versus poor reading attitudes, improving their reading attitudes and enhancing their reading achievement. However, extreme classmate differences in reading attitudes yielded lower student achievement, consistent with less homophily (Brechwald & Prinstein, 2011).

As with the above classmate past achievement and family SES results, the classmate reading attitude results suggest that a moderate mixtures of students with different reading attitudes might increase reading achievement overall. Specifically, assigning some students with positive reading attitudes into the same classes as low-achieving students can improve the reading achievement of the latter without harming that of the former.

Limitations and Future Studies

This study had five major limitations. First, these data do not indicate the source of schoolmates' and classmates' homogeneity of past achievement; it might stem from human decisions (streaming and tracking) or the natural homogeneity of students within a school district. Future studies can identify different sources of homogeneity and examine whether their consequences differ (e.g., overt tracking might produce labeling effects [Ferri & Connor, 2005] while natural homogeneity does not). Second, high costs and logistics limitations precluded collecting pretest scores of students before schooling in 40 countries, so parent reporting of their children's basic literacy skills (e.g., recognizing alphabet letters) before attending school was used (reliability = 0.95). Third, this study focused on students' reading performance only. Further studies can include different aspects of academic performance, such as mathematics or science, to provide a more comprehensive picture of school streaming and classroom tracking influences. Fourth, results in this study were correlational and did not warrant causal interpretations. Future studies can analyze longitudinal data to examine whether the above factors have causal relationships with student achievement. Lastly, we did not demonstrate mechanisms that drove the above results; instead, we provided evidence to reject some mechanisms and to show that others remain plausible. Future studies can gather additional data on teachers' and students' perceptions and behaviors to test these potential mechanisms.

Conclusion

This study investigated whether grouping students with similar past literacy skills, family SES, or reading attitude together within schools or classrooms was linked to their reading achievement to explore their possible mechanisms. Greater similarity of students' past literacy skills within a school (streaming) was linked to higher student reading achievement (customized instruction), although extremely homogeneous schools weakened this link (fewer help opportunities). However, when classmates' past literacy skills differed more (less classroom tracking), students had higher reading achievement (more help opportunities), although extreme differences weakened this link for high-achieving students (less customized instruction). The classmate past achievement results were also consistent with classmate sharing of ideas and help opportunities. Students had higher reading achievement when classmates had either stronger past literacy skills (sharing ideas) or extremely poor ones (help opportunities).

The results also suggest that classmates shared educational resources. When classmates had more educational resources at home or higher family SES, students had higher reading achievement (sharing educational resources), although with diminishing marginal returns for family SES. Also, when classmates' family SES differed more (more diversity), high-achieving students had higher reading achievement (Matthew effect).

Classmates' reading attitudes and their differences were also linked to higher student achievement. When classmates had better reading attitudes, low-achieving students had higher reading achievement (modeling, norm). Also, when classmates' reading attitudes differed more, students had higher reading achievement (contrasting cases), although extremely high differences weakened this link (homophily). By understanding how classmates and schoolmates influence students, educators can reallocate students into classrooms and schools to improve their academic achievement.

References

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (2003). *On the success of failure: A reassessment of the effects of retention in the primary school grades*. New York, NY: Cambridge University Press.

Ansalone, G., & Biafora, F. (2004). Elementary school teachers' perceptions and attitudes to the educational structure of tracking. *Education, 125*, 249–258.

Baker, D. P., Goesling, B., & Letendre, G. K. (2002). Socioeconomic status, school quality, and national economic development. *Comparative Education Review, 46,* 291–312.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory*. Boca Raton, FL: CRC.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology, 104,* 682–699. http://dx.doi.org/10.1037/a0027608

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika, 93,* 491–507. http://dx.doi.org/10.1093/biomet/93.3.491

Blatchford, P., Pellegrini, A. D., & Baines, E. (2016). *The child at school: Interactions with peers and teachers* (2nd ed.). New York, NY: Routledge.

Boaler, J. (2013). Ability and mathematics: The mindset revolution that is reshaping education. *Forum, 55,* I, 143–152.

Brechwald, W. A., & Prinstein, M. J. (2011). Beyond homophily: A decade of advances in understanding peer influence processes. *Journal of Research on Adolescence, 21,* 166–179. http://dx.doi.org/10.1111/j.1532-7795.2010.00721.x

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007). International surveys of educational achievement: How robust are the findings? *Journal of the Royal Statistical Society Series A: Statistics in Society, 170,* 623–646. http://dx.doi.org/10.1111/j.1467-985X.2006.00439.x

Caldas, S. J., & Bankston, C., III. (1997). Effect of school population socioeconomic status on individual academic achievement. *The Journal of Educational Research, 90,* 269–277. http://dx.doi.org/10.1080/00220671.1997.10544583

Chiu, M. M. (2007). Families, economies, cultures, and science achievement in 41 countries: Country-, school-, and student-level analyses. *Journal of Family Psychology, 21,* 510–519. http://dx.doi.org/10.1037/0893-3200.21.3.510

Chiu, M. M. (2008). Inequality mechanisms that hurt both privileged and disadvantaged students' learning. In I. H. Wadell (Ed.), *Income distribution: Inequalities, impacts and incentives* (pp. 79–98). Hauppauge, NY: Nova Science.

Chiu, M. M. (2010). Inequality, family, school, and mathematics achievement. *Social Forces, 88,* 1645–1676. http://dx.doi.org/10.1353/sof.2010.0019

Chiu, M. M. (2015). Family inequality, school inequalities and mathematics achievement in 65 countries: Microeconomic mechanisms of rent seeking and diminishing marginal returns. *Teachers College Record, 117,* 1–32.

Chiu, M. M., & Chow, B. W. Y. (2010). Culture, motivation, and reading achievement: High school students in 41 countries. *Learning and Individual Differences, 20,* 579–592. http://dx.doi.org/10.1016/j.lindif.2010.03.007

Chiu, M. M., & Chow, B. W. Y. (2015). Classmate characteristics and student achievement in 33 countries: Classmates' past achievement, family socioeconomic status, educational resources, and attitudes toward reading. *Journal of Educational Psychology, 107,* 152–169. http://dx.doi.org/10.1037/a0036897

Chiu, M. M., & Khoo, L. (2005). Effects of resources, inequality, and privilege bias on achievement: Country, school, and student level analyses. *American Educational Research Journal, 42,* 575–603. http://dx.doi.org/10.3102/00028312042004575

Chiu, M. M., & Klassen, R. M. (2009). Calibration of reading self-concept and reading achievement among 15-year-olds: Cultural differences in 34 countries. *Learning and Individual Differences, 19,* 372–386. http://dx.doi.org/10.1016/j.lindif.2008.10.004

Chiu, M. M., & McBride-Chang, C. (2006). Gender, context, and reading: A comparison of students in 41 countries. *Scientific Studies of Reading, 10,* 331–362. http://dx.doi.org/10.1207/s1532799xssr1004_1

Chiu, M. M., & McBride-Chang, C. (2010). Family and reading in 41 countries: Differences across cultures and students. *Scientific Studies of Reading, 14,* 514–543. http://dx.doi.org/10.1080/10888431003623520

Chiu, M. M., & Walker, A. D. (2007). Leadership for social justice in Hong Kong schools. *Journal of Educational Administration, 45,* 724–739. http://dx.doi.org/10.1108/09578230710829900

Chmielewski, A. K. (2014). An international comparison of achievement inequality in within- and between-school tracking systems. *American Journal of Education, 120,* 293–324. http://dx.doi.org/10.1086/675529

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students' mathematics self-concept. *American Educational Research Journal, 50,* 925–957. http://dx.doi.org/10.3102/0002831213489843

Davalos, D. B., Chavez, E. L., & Guardiola, R. J. (2005). Effects of perceived parental school support and family communication on delinquent behaviors in Latinos and White non-Latinos. *Cultural Diversity & Ethnic Minority Psychology, 11,* 57–68. http://dx.doi.org/10.1037/1099-9809.11.1.57

De Dreu, C. K. W., & West, M. A. (2001). Minority dissent and team innovation: The importance of participation in decision making. *Journal of Applied Psychology, 86,* 1191–1201. http://dx.doi.org/10.1037/0021-9010.86.6.1191

Ding, W., & Lehrer, S. F. (2007). Do peers affect student achievement in China's secondary schools? *The Review of Economics and Statistics, 89,* 300–312. http://dx.doi.org/10.1162/rest.89.2.300

Edmunds, K. M., & Bauserman, K. L. (2006). What teachers can learn about reading motivation through conversations with children. *The Reading Teacher, 59,* 414–424. http://dx.doi.org/10.1598/RT.59.5.1

Ferri, B. A., & Connor, D. J. (2005). Tools of exclusion: Race, disability, and (re)segregated education. *Teachers College Record, 107,* 453–474. http://dx.doi.org/10.1111/j.1467-9620.2005.00483.x

Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21,* 1–56.

Glogger, I., Holzapfel, L., Kappich, J., Schwonke, R., Nückles, M., & Renkl, A. (2013). Development and evaluation of a computer-based learning environment for teachers: Assessment of learning strategies in learning journals. *Education Research International, 2013,* 785065. http://dx.doi.org/10.1155/2013/785065

Goldstein, H. (2011). *Multilevel statistical models*. Sydney, New South Wales, Australia: Edward Arnold.

Guthrie, J. T., Klauda, S. L., & Morrison, D. A. (2012). Motivation, achievement, and classroom contexts for information book reading. In J. T. Guthrie, A. Wigfield & S. L. Klauda's (Eds.), *Adolescents' engagement in academic literacy*. College Park, MD: University of Maryland.

Guyon, N., Maurin, E., & McNally, S. (2012). The effect of tracking students by ability into different schools: A natural experiment. *The Journal of Human Resources, 47,* 684–721. http://dx.doi.org/10.3368/jhr.47.3.684

Hanson, B. A. (2002). *ICL*. Retrieved from http://www.b-a-h.com/software/irt/icl/

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics, 18,* 527–544. http://dx.doi.org/10.1002/jae.741

Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal, 116,* C63–C76. http://dx.doi.org/10.1111/j.1468-0297.2006.01076.x

Israel, G. D., Beaulieu, L. J., & Hartless, G. (2001). The influence of family and community social capital on educational achievement.

Rural Sociology, 66, 43–68. http://dx.doi.org/10.1111/j.1549-0831.2001.tb00054.x

Jakubowski, M., Patrinos, H. A., Porta, E. E., & Wiśniewski, J. (2016). The effects of delaying tracking in secondary school: Evidence from the 1999 education reform in Poland. *Education Economics, 24,* 557–572. http://dx.doi.org/10.1080/09645292.2016.1149548

Jansen, M., Schroeders, U., Lüdtke, O., & Marsh, H. W. (2015). Contrast and assimilation effects of dimensional comparisons in five subjects: An extension of the I/E model. *Journal of Educational Psychology, 107,* 1086–1101. http://dx.doi.org/10.1037/edu0000021

Johnson, D. W., & Johnson, R. (1999). *Learning together and alone: Cooperative, competitive, and individualistic learning.* Boston, MA: Allyn & Bacon.

Jöreskog, K., & Sörbom, D. (2012). *LISREL 9.1.* New York, NY: Scientific Software International.

Kaplan, S. N., Guzman, I., & Tomlinson, C. A. (2009). *Using the parallel curriculum model in urban settings, grades K–8.* Thousand Oaks, CA: Corwin.

Kennedy, P. (2008). *Guide to econometrics.* Hoboken, NJ: Wiley-Blackwell.

Kerr, S. P., Pekkarinen, T. P., & Uusitalo, R. (2013). School tracking and development of cognitive skills. *Journal of Labor Economics, 31,* 577–602. http://dx.doi.org/10.1086/669493

Kindermann, T. A. (2007). Effects of naturally existing peer groups on changes in academic engagement in a cohort of sixth graders. *Child Development, 78,* 1186–1203.

Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1,* 66–88. http://dx.doi.org/10.1080/19345740701692522

Lee, J. S., & Bowen, N. K. (2006). Parent involvement, cultural capital, and the achievement gap among elementary school children. *American Educational Research Journal, 43,* 193–218. http://dx.doi.org/10.3102/00028312043002193

Liu, W. C., Wang, C. K. J., & Parkins, E. J. (2005). A longitudinal study of students' academic self-concept in a streamed setting: The Singapore context. *British Journal of Educational Psychology, 75,* 567–586. http://dx.doi.org/10.1348/000709905X42239

Lleras, C., & Rangel, C. (2009). Ability grouping practices in elementary school and African American/Hispanic achievement. *American Journal of Education, 115,* 279–304. http://dx.doi.org/10.1086/595667

Lucas, S. R., & Berends, M. (2002). Sociodemographic diversity, correlated achievement, and de facto tracking. *Sociology of Education, 75,* 328–348. http://dx.doi.org/10.2307/3090282

Mankiw, N. G. (2014). *Principles of economics* (7th ed.). Boston, MA: Cengage Learning.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2003). *PIRLS 2001 Technical Report.* Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 Technical Report.* Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

May, H. (2006). A multilevel Bayesian IRT method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics, 31,* 63–79. http://dx.doi.org/10.3102/10769986031001063

McKenzie, D. J. (2005). Measuring inequality with asset indicators. *Journal of Population Economics, 18,* 229–260.

Mullis, I. V., Kennedy, A. M., Martin, M. O., & Sainsbury, M. (2006). *PIRLS 2006 Assessment framework and specifications: Progress in international reading literacy study.* Chestnut Hill, MA: PIRLS International Study Center.

Opdenakker, M.-C., & Van Damme, J. (2001). Relationship between school composition and characteristics of school process and their effect on mathematic achievement. *British Educational Research Journal, 27,* 407–432. http://dx.doi.org/10.1080/01411920120071434

Paulus, P. B., & Brown, V. (2003). Ideational creativity in groups. In P. B. Paulus & B. A. Nijstad (Eds.), *Group creativity* (pp. 110–136). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195147308.003.0006

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research. *Review of Educational Research, 74,* 525–556. http://dx.doi.org/10.3102/00346543074004525

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95,* 667–686. http://dx.doi.org/10.1037/0022-0663.95.4.667

Rasbash, J., Steele, F., Browne, W. J. and Goldstein, H. (2015). *MLwiN 2.33.* Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Rigney, D. (2013). *The Matthew effect: How advantage begets further advantage.* New York, NY: Columbia University Press.

Roelle, J., & Berthold, K. (2015). Effects of comparing contrasting cases on learning from subsequent explanations. *Cognition and Instruction, 33,* 199–225. http://dx.doi.org/10.1080/07370008.2015.1063636

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity. *Journal of the American Statistical Association, 96,* 20–31. http://dx.doi.org/10.1198/016214501750332668

Rowe, K. S., & Rowe, K. J. (1997). Norms for parental ratings on Conners' Abbreviated Parent–Teacher Questionnaire: Implications for the design of behavioral rating inventories and analyses of data derived from them. *Journal of Abnormal Child Psychology, 25,* 425–451. http://dx.doi.org/10.1023/A:1022678013979

Schippers, M. C., Den Hartog, D. N., & Koopman, P. L. (2007). Reflexivity in teams: A measure and correlates. *Applied Psychology, 56,* 189–211. http://dx.doi.org/10.1111/j.1464-0597.2006.00250.x

Schulz, W. (2003). *Validating questionnaire constructs in international studies.* Camberwell, Victoria, Australia: Australian Council for Educational Research.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16,* 475–522. http://dx.doi.org/10.1207/s1532690xci1604_4

Sharan, Y. (2010). Cooperative learning for academic and social gains: Valued pedagogy, problematic practice. *European Journal of Education, 45,* 300–313. http://dx.doi.org/10.1111/j.1465-3435.2010.01430.x

Skibbe, L. E., Phillips, B. M., Day, S. L., Brophy-Herb, H. E., & Connor, C. M. (2012). Children's early literacy growth in relation to classmates' self-regulation. *Journal of Educational Psychology, 104,* 541–553. http://dx.doi.org/10.1037/a0029153

Smith, B. J. (2013). Dropouts and differentiation: Toward understanding and prevention. In E. F. Sparapani (Ed.), *Differentiated instruction: Content area applications and other considerations for teaching in grades 5–12 in the twenty-first century* (pp. 221–230). Lanham, MD: University Press of America.

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research, 83,* 357–385. http://dx.doi.org/10.3102/0034654313483907

Tribushinina, E. (2008). *Cognitive reference points.* Utrecht, the Netherlands: Netherlands Graduate School of Linguistics.

van Knippenberg, D., De Dreu, C. K., & Homan, A. C. (2004). Work group diversity and group performance: An integrative model and research agenda. *Journal of Applied Psychology, 89,* 1008–1022. http://dx.doi.org/10.1037/0021-9010.89.6.1008

van Knippenberg, D., & Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology, 58,* 515–541. http://dx.doi.org/10.1146/annurev.psych.58.110405.085546

van Offenbeek, M. (2001). Processes and outcomes of team learning. *European Journal of Work and Organizational Psychology, 10,* 303–317. http://dx.doi.org/10.1080/13594320143000690

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450.

Watanabe, M. (2008). Tracking in the era of high stakes state accountability reform: Case studies of classroom instruction in North Carolina. *Teachers College Record, 110,* 489–534.

Watson, W. E., Johnson, L., & Zgourides, G. D. (2002). The influence of ethnic diversity on leadership, group process, and performance: An examination of learning teams. *International Journal of Intercultural Relations, 26,* 1–16. http://dx.doi.org/10.1016/S0147-1767(01)00032-3

Westwood, P. (2013). *Inclusive and adaptive teaching: Meeting the challenge of diversity in the classroom.* New York, NY: Routledge.

Willms, J. D. (1999). Quality and inequality in children's literacy: The effects of families, schools, and communities. In D. P. Keating & C. Hertzman (Eds.), *Developmental health and the wealth of nations* (pp. 72–93). New York, NY: Guilford Press.

Winters, M. A., Haight, R. C., Swaim, T. T., & Pickering, K. A. (2013). The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data. *Economics of Education Review, 34,* 69–75. http://dx.doi.org/10.1016/j.econedurev.2013.01.007

World Bank. (2007). *The world development report 2006.* New York, NY: Oxford University Press.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31,* 114–128. http://dx.doi.org/10.1016/j.stueduc.2005.05.005

Appendix A

Statistical Power for Effect Sizes at Each Level

	Effect size			
Level	1	2	3	4
3) Country	.07	.13	.23	.37
2) School	.43	.94	1.00	1.00
1) Student	.44	.95	1.00	1.00

(*Appendices continue*)

Appendix B

Correlation-Variance-Covariance Matrix of Outcome Variables and Explanatory Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	**12807**	8.20	20.44	−.415	−1.39	36.74	19.29	53.09	−3.97	26.93	42.96	−21.83	6.26	−19.56	4.06	13.73	35.11	.90	.91	5.69	7.43
2	.07	**1.00**	.01	1.44	.00	.08	.10	.06	.23	.03	.00	.02	.00	.05	.05	.10	.15	.02	−.03	.01	.00
3	.28	.01	**.43**	−1.73	−.04	.15	.03	.27	.01	.15	.27	.01	.02	−.15	.00	−.09	.04	−.09	−.03	.02	.06
4	−.46	.18	−.33	**62.46**	.09	−1.18	−.54	−2.46	1.44	−1.18	−2.46	1.35	−.43	1.67	.02	.23	−.83	.23	−.36	−.31	−.36
5	−.13	−.03	−.65	.12	**.01**	−.02	−.01	−.03	.00	−.02	−.03	.00	.00	.02	.00	.01	.00	.01	.00	.00	−.01
6	.33	.08	.24	−.15	−.21	**1.00**	.28	.49	.03	.37	.32	−.05	.03	−.09	.00	.02	.15	−.03	.01	.05	.03
7	.17	.10	.05	−.07	−.07	.28	**1.00**	.30	.02	.13	.13	−.04	.01	−.03	.01	.10	.13	.01	.01	.02	.01
8	.47	.06	.41	−.31	−.32	.49	.30	**1.00**	.00	.32	.46	−.05	.04	−.17	.02	.06	.19	−.04	.00	.04	.05
9	−.07	.48	.02	.38	−.07	.06	.04	.01	**.23**	.03	.00	.04	.00	.05	.01	.02	.01	.02	−.03	.01	.00
10	.39	.05	.39	−.25	−.35	.61	.21	.52	.10	**.37**	.32	−.04	.03	−.09	.00	−.03	.06	−.03	.01	.05	.03
11	.56	.01	.60	−.46	−.47	.46	.19	.68	.01	.76	**.46**	−.06	.04	−.17	.00	−.04	.08	−.04	.00	.04	.05
12	−.19	.02	.01	.16	−.01	−.05	−.04	−.05	.08	−.06	−.08	**1**	−.02	.05	−.03	−.06	−.10	−.03	−.03	−.02	.02
13	.15	−.01	.07	−.15	−.10	.09	.03	.11	−.03	.16	.15	−.06	**.13**	−.02	.01	.00	.01	.00	.01	.00	.01
14	−.28	.08	−.39	.35	.29	−.15	−.05	−.27	.16	−.25	−.40	.08	−.10	**.37**	.00	.05	−.03	.05	.00	−.01	−.03
15	.07	.09	.00	.01	−.01	.00	.01	.03	.03	.01	.01	−.06	.04	.00	**.25**	.09	.03	.01	.00	.00	.00
16	.12	.10	−.14	.03	.14	.02	.10	.06	.04	−.04	−.06	−.06	.01	.07	.18	**1.00**	.20	.17	.02	.01	−.06
17	.31	.15	.06	−.10	−.01	.15	.13	.19	.03	.10	.11	−.10	.04	−.05	.05	.20	**1.00**	.02	.02	.02	.01
18	.02	.04	−.34	.07	.33	−.07	.03	−.11	.09	−.11	−.16	−.07	.02	.18	.05	.41	.06	**.17**	.02	.01	−.06
19	.02	−.07	−.11	−.13	.08	.03	.03	−.01	−.15	.05	−.01	−.08	.06	−.02	.00	.06	.05	.14	**.12**	.02	.00
20	.16	.04	.08	−.13	−.07	.16	.07	.14	.08	.26	.21	−.05	.03	−.07	.01	.03	.06	.07	.16	**.10**	.00
21	.19	−.01	.27	−.13	−.21	.08	.02	.16	−.03	.13	.23	.05	.08	−.14	−.01	−.18	.04	−.43	−.03	.00	**.12**

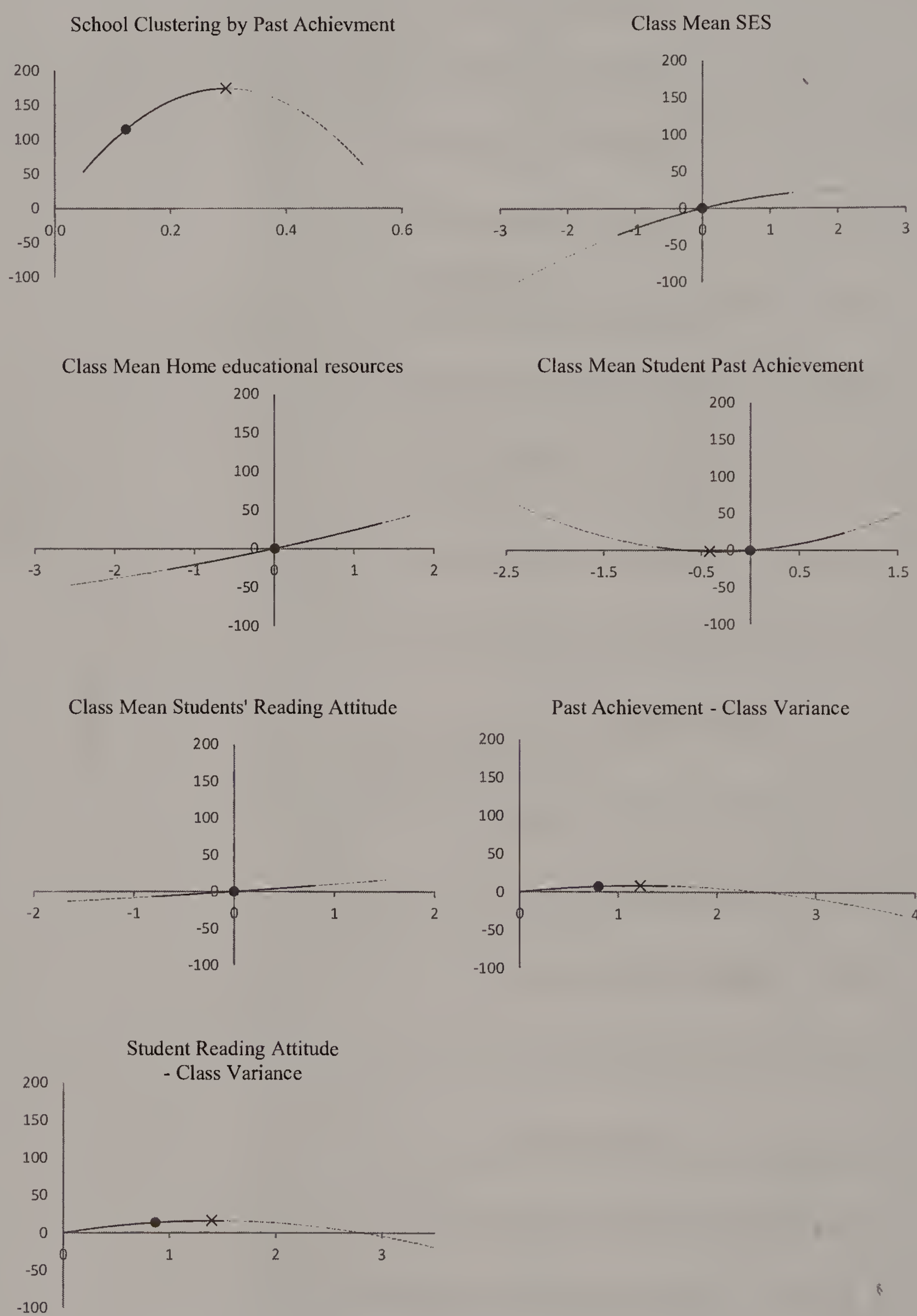
Note. The correlations, variances, and co-variances are along the lower left triangle, diagonal, and upper right triangle of the matrix. 1) Reading achievement, 2) Past literacy skills, 3) Log GDP per capita, 4) GINI, 5) Clustering of students by past literacy skills, 6) Family SES, 7) Parents' attitude towards reading, 8) Home educational resources, 9) Class mean past literacy skills, 10) Class mean SES, 11) Class mean home educational resources, 12) School violence, 13) Female teacher, 14) HW mismatch, 15) Girl, 16) Students' attitude towards reading, 17) Students' reading self-concept reading, 18) Class mean students' attitude towards reading, 19) Variance of classmates' past literacy skills, 20) Variance of classmate SES, 21) Variance of classmates' attitude towards reading.

(*Appendices continue*)

Appendix C

Non-Linear Relations of Explanatory Variables With Reading Achievement

All graphs have the same reading achievement y-axis scale of −100 to 200. Each blue dot indicates an explanatory variable's mean value and the solid blue curve indicates values within two standard deviations of this mean. The dashed curves indicate extreme values outside two standard deviations of this mean. Some curves have a turning point, indicated by a red X (minimum or maximum).



See the online article for the color version of this figure.

Received July 26, 2016
Revision received December 13, 2016
Accepted December 13, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 7, 935–955

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000186

Framework for Disciplinary Writing in Science Grades 6–12: A National Survey

Sally Valentino Drew
Central Connecticut State University

Natalie G. Olinghouse
University of Connecticut

Michael Faggella-Luby
Texas Christian University

Megan E. Welsh
University of California, Davis

This study investigated the current state of writing instruction in science classes (Grades 6–12). A random sample of certified science teachers from the United States (N = 287) was electronically surveyed. Participants reported on their purposes for teaching writing, the writing assignments most often given to students, use of evidence-based writing practices, and the instructional adaptations made for struggling writers in science class. Typical practice was examined against a theoretical framework for disciplinary writing in science that articulates research-based and evidence-based practices to improve writing and learning outcomes for all students, including struggling writers. Descriptive results, exploratory factor analysis, and examination of differences between groups (middle school and high school teachers) revealed concerns about the quantity and quality of writing practices in secondary science classrooms, especially at the high school level. Although the majority of participants report to include writing purposely to accompany the inquiry process, most of the writing tasks teachers report to include in science require minimal composition. Participants report to include evidence-based practices for teaching writing and adapting instruction to support struggling writers at a frequency range of once per year to once per quarter. Results inform recommendations for teacher education, professional development, and instructional reform for disciplinary writing in science that supports all learners.

Keywords: disciplinary literacy, evidence-based writing practices, science education, struggling writers, writing instruction

Supplemental materials: http://dx.doi.org/10.1037/edu0000186.supp

Science education reform is focused on infusing the way scientists think and work into daily instruction (National Research Council [NRC], 2012), rather than having students learn about science as a bystander. A core component of the work of scientists is communicating scientific understanding in writing. In fact, writing is a unifying element across all eight of the science and engineering practices highlighted in the Next Generation Science Standards (Achieve, 2013), and is directly addressed in three of the eight practices (constructing explanations; engaging in argument from evidence; and obtaining, evaluating, and communicating information).

Despite an increased demand for writing in science, many adolescents struggle to effectively convey thoughts and ideas in writing in science or other content areas. Recently released data show nearly 75% of eighth and twelfth graders are performing at or below the basic level on the most recent National Assessment of Educational Progress (NAEP) writing assessment (National Center for Education Statistics [NCES], 2012). Furthermore, unique writing demands within each academic discipline (math, science, history, literature) exacerbate this crisis. Middle and high school students are not only expected to meet general reading and writing competencies, but also to master literacy skills and strategies in order to unlock and convey content knowledge within a given discipline, such as science (NRC, 2012). Yet, similar numbers of students struggle with reading and writing performance in academic content areas (Fang & Coatoam, 2013).

The Science NAEP requires students to use writing to predict, describe, explain, and draw conclusions about various science topics (NCES, 2012). For example, on a sample item that asked students to draw a conclusion from a given data table and construct

This article was published Online First March 13, 2017.

Sally Valentino Drew, Department of Special Education and Interventions, Central Connecticut State University; Natalie G. Olinghouse, Department of Educational Psychology, University of Connecticut; Michael Faggella-Luby, College of Education ANSERS Institute, Texas Christian University; Megan E. Welsh, School of Education, University of California, Davis.

This paper shares the results of a dissertation study that was conducted while Sally Valentino Drew was a doctoral student at the University of Connecticut, with Natalie G. Olinghouse and Michael Faggella-Luby as dissertation advisors.

Correspondence concerning this article should be addressed to Sally Valentino Drew, Department of Special Education, and Interventions Central Connecticut State University, 1615 Stanley Street, New Britain, CT 06050. E-mail: drewsav@ccsu.edu

a written response explaining their answer, only 15% of respondents provided a complete answer. Overall, adolescents are doing slightly better nationally on the science NAEP (65% of eighth graders are performing at or below basic), yet the data are not disaggregated for question type, and therefore conclusions cannot be drawn about 'writing in science' items specifically. Given the finding reported for one constructed response item (only 15% of respondents providing a complete answer), the results from writing items alone might depict even greater struggle.

Context of Writing Instruction in Science Classrooms

As such, adolescents are struggling to meet minimum writing competencies within and beyond science. However, research has shown that quality instruction can influence students' performance in writing (see Graham & Perin, 2007a, 2007b, and 2007c for meta-analyses of adolescent writing research), especially with populations of struggling writers. Meta-analyses reveal that evidence-based practices[1] such as writing strategy instruction (effect size: Cohen's $d = 0.82$) and process writing (effect size: Cohen's $d = 0.32$) have been shown to strongly and consistently improve adolescents' writing quality across content areas (Graham & Perin, 2007c). However, many of these practices are not making their way into content classrooms, including science (Applebee & Langer, 2011; Gillespie, Graham, Kiuhara, & Hebert, 2014; Graham, Capizzi, Harris, Hebert, & Morphy, 2014; Kiuhara, Graham, & Hawken, 2009). According to study results, there are many possible reasons why evidence-based writing practices are not being implemented in secondary classrooms.

First, many content teachers do not report to feel confident teaching writing. Science teachers, specifically, report feeling less prepared to teach writing than do language arts and social studies teachers (Gillespie et al., 2014; Kiuhara et al., 2009). With the exception of math teachers, science teachers spend less time than those in other content areas teaching extended writing of longer than a paragraph (Applebee & Langer, 2011).

Second, science teachers may be unaware of the most effective ways to teach writing. Writing at the secondary level—as with reading—draws upon foundational skills and strategies, but is also unique within each of the content areas. General writing skills and strategies apply across content areas, but there are also specialized—or discipline-specific—approaches, skills, strategies, routines, audiences, and purposes for writing that are different in science than they are in language arts or social studies (Shanahan & Shanahan, 2008). Research is just beginning to emerge to articulate how content teachers can use discipline-specific writing instruction in content area classrooms or to adapt general writing strategies for discipline-specific learning (Faggella-Luby, Graner, Deshler, & Drew, 2012). Yet, survey data (Applebee & Langer, 2011; Gillespie et al., 2014; Graham et al., 2014; Kiuhara et al., 2009) illustrate that neither discipline-specific nor general writing strategies are making their way into science classrooms regularly.

A third reason why teachers may not be teaching writing in science is that they may see extended writing as an add-on to their content instruction, and not integral to promoting learning within their content. When teachers see science content and writing as separate, they are more likely to privilege science content, as it is their primary responsibility. Science teachers may not yet see the value in using extended writing to deepen thinking and learning, and to enable participation in the scientific learning community, as is highlighted in the new national science standards (Achieve, 2013). Per writing meta-analyses results (Graham & Perin, 2007c), writing to learn tasks have shown slight to moderate effect sizes across writing studies (average $d = 0.23$). However, the majority of writing that science teachers assign includes completion of worksheets and step-by-step directions (Kiuhara et al., 2009) or tasks that do not require much analysis and interpretation (Gillespie et al., 2014), which may be less likely to promote learning. These tasks demonstrate a *restricted* (Applebee, 1981, 1984) view of writing, in which tasks are noncompositional and teacher-directed, emphasizing the transaction of knowledge demonstrated from the student for the teacher.

The fourth reason may be the tension teachers feel in choosing between research-based inquiry pedagogy for science instruction and evidence-based pedagogy of explicit instruction for writing development. Science educator preparation and in-service workshops focus on inquiry as the preferred pedagogy. Inquiry pedagogy emerges from a constructivist-learning paradigm that models the work of practicing scientists (NRC, 1996) and promotes deep learning of science concepts through enculturation (Moje, 2015).

Scientific inquiry refers to the diverse ways in which scientists study the natural world and propose explanations based on the evidence derived from their work. Inquiry also refers to the activities of students in which they develop knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world. (NRC, 1996, p. 23)

With inquiry approaches, students direct more of their own learning, whereas evidence-based pedagogies for struggling writers rely on explicit instruction based on a cognitive and behavioral learning perspective (Archer & Hughes, 2011). Inquiry writing tasks have shown notable effect sizes ($d = 0.32$) across studies (Graham & Perin, 2007c), yet teachers may not yet be aware of how these two pedagogical practices can coexist effectively. Research is emerging to demonstrate that explicit instruction of writing skills and strategies can be embedded within the context of inquiry-based science (Cervetti, Barber, Dorph, Pearson, & Goldschmidt, 2012; McNeill, 2011; McNeill & Krajcik, 2009).

Given the recent writing NAEP scores of nearly 75% of eighth and twelfth graders performing at or below basic (NCES, 2012), and the report that 81% of students with disabilities are enrolled in general education classrooms for between 40 and 79% of their day (NCES, 2016), science teachers' classrooms most likely include struggling writers. Teachers may not be teaching writing in science because they are not sure how to adapt instruction for their large number of struggling writers. In Kiuhara and colleagues' survey study (2009), science teachers were reported to utilize the least

[1] The authors adhere to the description and distinction of evidence-based practices and research-based practices as defined by Cook and colleagues (2008, 2013). Evidence-based practices are instructional approaches shown by a substantial body of rigorous scientific research as most likely to meaningfully improve student outcomes (Cook, Tankersley, Cook, & Landrum, 2008), whereas instructional techniques that are termed research-based are supported by a growing body of literature as likely to improve student outcomes, but currently lack the quantity and quality of studies to demonstrate persistent effects for all learners (Cook & Cook, 2013). This study draws upon the evidence and research bases that articulate support for writing in science practices.

number of writing instructional adaptations, and used them the least frequently of the content areas surveyed.

Theoretical Framework

Given the problem of research- and evidence-based writing practices not making their way into science classrooms regularly, this paper introduces the *Framework for Disciplinary Writing in Science* (Framework) as a research-based theoretical model for enhancing writing instruction to meet the needs of all students within a science classroom. The survey study reported in this paper examines the alignment of the Framework to typical practice as reported by secondary science teachers.

The Framework considers and extends two primary research-based models for writing to learn in science that support teacher decision-making about the inclusion of writing in science classes. Prain and Hand (1996; elaborated on in Hand & Prain, 2002) outline core aspects of writing to learn as five "keys" (purpose, genre, audience, topic, and method of production) to unlocking high quality writing tasks that deepen learning in secondary science class. Hand and colleagues' model (1996, 2002) clarifies how teachers can diversify the learning and assessment experiences in science class through exposing students to diverse genres.

Klein and colleagues' Framework for Content Area Writing (Klein & Kirkpatrick, 2010; Klein & Rose, 2010) offers a set of design principles teachers can use to create writing intensive content units. Klein and colleagues provide recommendations for a comprehensive content-focused writing program including: writing routines, teaching strategies, specified genres, motivational features, and assessment practices that promote writing to learn. Analytic genres such as explanation and argumentation promote student thinking and synthesis; these genres are emphasized within the model as most likely to promote learning. Klein and colleagues aptly embed cognitive writing strategy instruction (Graham, 2006) within inquiry cycles as a synergistic blend of both evidence-based practice in writing and inquiry-based models of instruction familiar to science teachers.

Theory of Disciplinary Writing Progressions

Building from the models of writing to learn described above, this study's Framework incorporates a polytheoretical approach to understanding the role of writing in science classes to build and extend knowledge and communicate understanding. The theory behind the Framework considers how students progress in writing skill and development throughout the grades in order to reach skilled science writing at the secondary level, where students use writing not only to demonstrate understanding of scientific concepts, but also to powerfully communicate their scientific theories and to build knowledge within a scientific learning community.

The model of disciplinary literacy articulated by the Shanahans (Shanahan & Shanahan, 2008, 2012) illustrates literacy development over time and proposes teaching foundational literacy skills and strategies at the elementary level, moving to teaching intermediate literacy skills and strategies in the middle grades, and finally, teaching discipline-specific literacy skills and strategies at the high school level and beyond (Shanahan & Shanahan, 2008). The goal of disciplinary literacy is for students to be prepared for sophisticated college and career literacy demands through instruction in discipline-specific literacy strategies within the content areas of mathematics, science, history/social studies, and literature at the secondary and postsecondary level (Fang & Schleppegrell, 2010; Shanahan & Shanahan, 2008).

The current disciplinary literacy model is reading-centric. Writing has its own unique stages of development and progressions (National Commission on Writing in America's Schools and Colleges, 2003), which have not yet been reflected in the Shanahan and Shanahan (2008, 2012) model. Building from the work of Hand and colleagues (Hand & Prain, 2002; Prain & Hand, 1996) and Klein and colleagues described above (Klein & Kirkpatrick, 2010; Klein & Rose, 2010), this study's polytheoretical Framework utilizes the organizational structure and some language from the Shanahan's disciplinary literacy model, but specifically articulates a theory of *writing* progressions based on Bereiter and Scardamalia's (1987) model of skilled writing. Within the Framework, skilled writing results in students being able to communicate key scientific ideas clearly and powerfully, building from a writing progression in which foundational composition is taught at the elementary level, intermediate composition is taught at the middle school level linking foundational approaches to discipline-specific approaches, and disciplinary composition is then taught at the high school level and beyond.

Foundational Composition

The goal of foundational composition is to learn how to write in science. This includes developing the habits of writing across content areas including science. These *general* skills and strategies are consistent across any writing content including, but not limited to skills and strategies for: handwriting, spelling, vocabulary, sentence construction, and paragraph writing (Santangelo & Olinghouse, 2009). Because young students are faced with a variety of challenges with content writing (reading to learn about content, transcribing ideas into text, accessing limited background knowledge, lacking discourse knowledge), a primary goal is for students to produce text with increased accuracy, speed, and length (fluency). For young students to be able to do this, they often use a 'knowledge telling' approach in which they record ideas about a topic as they come into their mind in a manner similar to stream of consciousness (Bereiter & Scardamalia, 1987). With 'knowledge telling' the only constraints the young writer faces are a focus on the topic, what they know about the topic, and what they have already written about their topic on paper. At the foundational composition stage, the young writer is working to develop handwriting, spelling, vocabulary, and text construction skills to fully express what he or she knows about the topic. It is more difficult for the elementary student to learn during the writing process because so much of the cognitive load is devoted to demands of writing production.

Intermediate Composition

The goal of intermediate composition is to transition from learning how to write in a content area such as science toward deepening learning through the writing process, or writing to learn (Bangert-Drowns, Hurley, & Wilkinson, 2004). In this phase, students acquire more sophisticated techniques for writing in science including how to use the writing process as a cognitive

process to solve a writing problem and to set goals, brainstorm, draft, revise, and edit texts. Writers in this stage learn how to self-regulate the use of the steps in the writing process as well as how to give and receive feedback. As students have developed writing fluency in the foundational composition phase, they are able to plan for the organization and composition of multiparagraph text. Intermediate composition utilizes students' foundational skills and strategies, and moves them into more advanced *general* writing techniques that are applicable across content areas. Intermediate composition leads students into discipline-specific writing skills and strategies in specific content areas, like science. Students begin to learn to use writing to deepen thinking and to refine scientific understanding.

The key outcome in intermediate composition is for students to learn that writing is a problem solving process. Each writer is faced with certain constraints—the topic, the task, the audience, the purpose for writing, the preferred genre, the organizational structure, the language to communicate meaning, as well as transcription skills. At the intermediate level, students have more control over word-level skills (spelling, handwriting, vocabulary), and therefore they are able to expend more energy to compose text purposefully. Bereiter and Scardamalia (1987) describe the problem-solving approach to writing at this stage as 'knowledge transforming'—the goal of skilled writing. Writers expend significant effort to compose text in a 'knowledge transforming' mode because they are cycling back and forth between two problem spaces: content and rhetoric. The content space requires writers to consider what they know about their topic, and what information will be most relevant to communicate. The rhetorical space requires writers to consider how they will most effectively communicate their understanding to their audience. The 'knowledge telling' process involved in foundational composition does not necessarily affect the writer's knowledge, whereas the 'knowledge transforming' process solidifies and expands the writer's knowledge (Bereiter & Scardamalia, 1987). By considering not only what the young writer knows about a topic, but also how to best communicate this knowledge, the writer's knowledge is refined and deepened.

Disciplinary Composition

The goal of disciplinary composition is to be able to use advanced compositional skills and strategies to communicate disciplinary knowledge to a particular audience. Students use the tools they developed through the foundational and intermediate composition phases to learn specific writing techniques to communicate disciplinary (science) understanding.

Scardamalia and Bereiter (1991, 2006, 2010; Bereiter & Scardamalia, 2010) offer a third approach to writing that applies within the disciplinary composition phase: knowledge building. 'Knowledge building' emerges from 'knowledge transforming' and represents the ultimate goal of writing—for the writer to transform a reader's understanding, while building up the knowledge base of a learning community. Within 'knowledge building' students actively engage in intentional and purposeful writing practices that benefit a learning community. Writing within the discipline of science is the avenue for expressing scientific understanding in the form of theories that explain and predict phenomena (Chuy et al., 2010). Thus, writing and science are inextricably linked in purpose. Deep learning is synonymous with 'knowledge building,' and it is how knowledge advances in the discipline (Bereiter & Scardamalia, 2010). Disciplinary composition through 'knowledge building' asks students to perform at the highest level of scientific literacy, beyond fact recall and procedural testing skills toward theory building (Chuy et al., 2010). Students ask and answer questions to compose and evaluate explanations about natural phenomena and scientific arguments based on evidence (NRC, 1996). Advanced writing that is performed at the 'knowledge building' level intentionally improves the learning of the writer beyond that explained in the 'knowledge transforming' model because it involves more than reorganizing and selecting knowledge to fit content and rhetorical constraints. The writer also considers a third space of how and where this knowledge fits within a scientific learning community (classroom, school, or greater scientific community). Writing to learn with 'knowledge building' affects the writer's knowledge, but its ultimate goal is to influence the knowledge base within a scientific learning community.

Progressing Along the Continuum

The progression for instruction based on this Framework considering typical writing development includes foundational composition taught in the early elementary grades (approximately K–2), intermediate composition taught at upper elementary and middle grades (approximately 3–6), and disciplinary composition taught at upper middle into high school and beyond (approximately Grades 7 and above). Yet, the boundaries between each phase are permeable based on students' developmental writing needs and curricular expectations. Teachers can utilize their understanding of the writing process and pedagogical content knowledge and skill to help each student develop along the continuum. For example, if struggling adolescent writers are lacking certain foundational composition skills and strategies, they can be taught those competencies in an intervention block or in small group instruction within the classroom setting, and still receive disciplinary composition instruction with their peers in science class.

Research-Based Components of the Framework for Disciplinary Writing in Science

There are four components that represent the purview and intention of the Framework to support teachers in making theoretically sound, research-based decisions about their purposes for writing in science, the assignments they choose, the instruction they provide to students, and the adaptations they make for struggling writers. Each element of the Framework emerged from an extensive literature review (Drew, 2013), and was informed by results of national survey studies of writing instruction that included science teachers (Applebee & Langer, 2011; Gillespie et al., 2014; Kiuhara et al., 2009). Figure 1 displays research- and evidence-based elements to support science teachers' decision-making with the goal of promoting student thinking, learning, and communicating in science. Elements of an effective writing program in science that are most supported by the Framework are underlined in Figure 1.

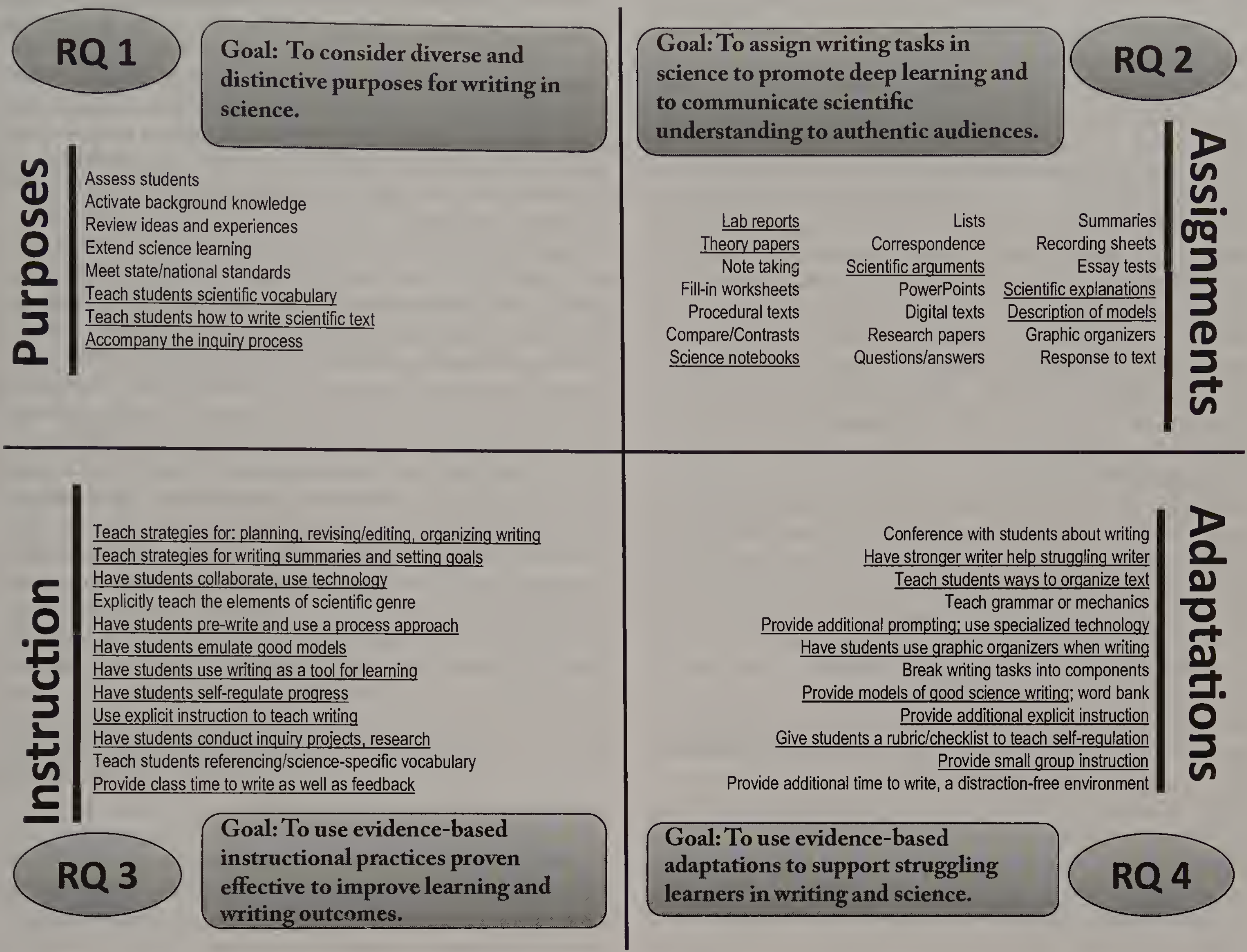


Figure 1. Guide for teacher decision making for writing in science. Research- and evidence-based elements of a writing program to support science teachers' decision-making with the goal of promoting student thinking, learning, and communicating in science. Underlined aspects have the strongest research and theoretical support to promote the goals of the Framework.

Consider Diverse and Distinctive Purposes for Writing in Science

In an effective writing program in science class, science teachers perceive writing as an essential element of what it means to do science (NRC, 1996, 2012). The purpose for writing in science becomes infused with the overarching goals of science education. Writing and science are synergistic (Pearson, Moje, & Greenleaf, 2010): writing is the vehicle for explaining scientific theories and the glue that connects those theories with the supporting procedures, data, claims, and evidence (Keys, Hand, Prain, & Collins, 1999). An essential purpose for writing in science is to accompany the inquiry process (Cervetti et al., 2012; Gaskins et al., 1994).

When students understand the nature of science, they can use writing to communicate scientific understanding and explanations of natural phenomena through scientific theory building. The purpose of the discipline of science is to contribute new knowledge to the scientific and broader community in the form of theories that explain and predict phenomena (NRC, 2012; Chuy et al., 2010; Scardamalia & Bereiter, 2010). To do this, scientists use a set of discipline-specific practices to guide theory development, reasoning, testing, and the communication of these understandings to build knowledge within a scientific learning community. The purpose of science education is to instill in students an understanding of science, so that they may become careful consumers of scientific information, and if they choose, possess the skills and knowledge to enter careers in science, engineering, and technology (NRC, 2012). Being able to write scientific text proficiently is an essential part of these goals.

Assign Writing Tasks in Science to Promote Deep Learning and to Communicate Scientific Understanding to Authentic Audiences

The Framework can help teachers select writing assignments in science that lend themselves toward writing to learn, and

ultimately to knowledge building. Writing is used within the disciplines to deepen content area learning (Bangert-Drowns, Hurley, & Wilkinson, 2004; Britton, 1970; Langer & Applebee, 2007; Newell, 2006). As writing and learning share an active and self-regulated feedback process of constructing meaning from prior knowledge and experience (Emig, 1977), "good writing and careful thinking go hand in hand" (Langer & Applebee, 2007, p. 3). Yet, not all writing tasks are equal in their ability to encourage thinking and learning. For the purposes of this paper, three different distinctions of tasks are made. First, the extent and length of student composition is addressed with the distinction between restricted writing tasks and analytic writing tasks (Applebee, 1981, 1984). With an approach aligned to the Framework, teachers would minimize restricted (teacher-controlled) tasks for prewriting or assessment purposes in science, and emphasize analytic (extended student-driven composition) tasks to best promote learning and knowledge building.

Applebee (1981, 1984) describes school-based writing functions within the content areas in the three categories of restricted writing, summary writing, and analytic writing. Within the literature on science writing, summary writing is considered more of a restricted writing task since the teacher controls much of the form and function. Therefore, the remainder of the tasks in science writing can fall within the categories of restricted or analytic tasks. Analytic writing tasks require a "complex manipulation of ideas" in a writer's attempt to make a claim and support it by selecting the most appropriate language (Newell, 2006, p. 239). Analytic writing before, during, and after scientific explorations serves as a record of learning, yet more importantly promotes self-regulation of scientific processes (Hand & Prain, 2002; Prain & Hand, 1996). Analytic writing practices include the written expression of thinking, often in the form of scientific explanation-arguments backed with claims, evidence, and reasoning (NRC, 2012). Analytic tasks require students to write extended compositions to share their thinking.

Restricted writing includes tasks in which teachers direct most of the writing, and students are not required to compose extended text (Applebee, 1981, 1984). Because writing transforms and builds learning through the process of composition (Bereiter & Scardamalia, 1987; Scardamalia & Bereiter, 2010), restricted writing tasks do not lend themselves to the same depth of learning as analytic writing. Yet, there is literature supporting the inclusion of these restricted writing practices, often preceding extended writing tasks. Examples of restricted writing tasks include: note taking, fill in worksheets, lists, and questions/short answers. Science teachers report to more regularly assign restricted writing tasks than other content area teachers except for math teachers (Applebee & Langer, 2011; Kiuhara et al., 2009).

Second, the purpose, audience, and context of the writing task can be explained by the difference between tasks that primarily promote learning for the writer (internal audience) and tasks that communicate to a broader audience (external audience). Borrowing Bereiter and Scardamalia's (1987) terminology to describe the relationship of the task to the purpose and audience, tasks that promote learning primarily for the writer (internal audience) have a 'knowledge transforming' focus. These tasks help students use writing to learn science content (taking notes, filling out graphic organizers, answering open-ended questions), for the purpose of deepening the writer's conceptual understanding. Although some of these tasks may eventually build toward compositions for a broader audience and purpose, the primary audience for such a task is the student writer. This contrasts with tasks intended to communicate to a broader (external) audience. These tasks align to Scardamalia and Bereiter's (2006, 2010) knowledge building theory. Although such tasks may begin as writing to learn, the ultimate goal of the task is to communicate findings to a broader audience (describe scientific models, communicate theories and scientific arguments, share research).

Third, the specificity of the task for science class can be examined by the distinction between general and discipline-specific tasks. Writing tasks or assignments that are not applicable to any content area other than science are deemed *discipline-specific* to science. Classroom practice will most likely begin to catch up to the emerging research in disciplinary literacy, but as of now, general writing practices have a more robust evidence base supporting student writing outcomes than disciplinary writing practices (Graham & Perin, 2007a, 2007b, 2007c). See Table 1 for examples of research-based discipline-specific writing tasks in science.

Table 1
Examples of Research-Based Discipline-Specific Writing Tasks in Science

Category	Definition
Description of models/concept maps	An explanation of a visual representation of a scientific concept or process. Typically models depict relationships among variables to explain phenomena. (Starr & Krajcik, 2013)
Lab reports	Description of process and findings from a scientific experiment, following the scientific method. Usually includes a section on introduction, procedures/materials, results, discussion, and conclusion. Exact format differs by teacher and grade level. Typically used by classroom teachers to assess learning. (Keys et al., 1999)
Scientific arguments	An essential scientific writing practice used to posit a set of reasons to explain scientific phenomena with ample backing (data, evidence). Crosscuts with explanation. (McNeill & Krajcik, 2009; McNeill, 2011)
Scientific explanations	An essential scientific practice used to write about natural phenomena by elaborating a series of evidence-based statements. Crosscuts with argument. (Keys et al., 1999; McNeill & Krajcik, 2009; McNeill, 2011)
Science notebooks	A writing journal used specifically in science to track the scientific process. Can be used to house a variety of types and forms of writing: to record the collection of data and observation of phenomena, to demonstrate analysis and synthesis of data, and to present arguments based on the data (Baxter et al., 2001).
Scientific theory papers	Writing that explains or builds the case for a scientific theory (Chuy et al., 2010; NRC, 2012). Typically provides reasons and evidence to support a description of the causes of natural phenomena.

Use Evidence-Based Writing Practices Proven Effective to Improve Writing and Learning Outcomes

Ensuring all students' success on the breadth and depth of writing assignments articulated in the Framework requires high quality writing instruction. Research indicates that quality instruction influences students' writing performance (Graham & Perin, 2007a, 2007b, 2007c). Evidence-based practices such as writing strategy instruction and process writing[2] have been shown to strongly and consistently improve adolescents' writing quality across content areas (Graham & Perin, 2007a, 2007b, 2007c). This is particularly important for struggling adolescent learners who may lack foundational writing skills and strategies, which can impede their ability to meet the demanding writing tasks outlined in the Framework.

Teaching writing is equally as important as assigning writing in science. Teachers can provide instruction to help students best communicate scientific understanding in writing so that students learn to write scientific genre effectively. Graham and Perin (2007c) outline evidence-based practices that are supported by rigorous research to help all adolescent students learn to write well and use writing as a tool for learning. Unfortunately, previous studies indicate that these practices are not consistently making their way into science classrooms (Applebee & Langer, 2011; Gillespie et al., 2014; Graham et al., 2014; Kiuhara et al., 2009).

Use Evidence-Based Adaptations to Support Struggling Learners in Writing and Science

Most of the studies investigating disciplinary writing practices in science class do not specifically include populations of struggling writers and students with disabilities (Drew, 2013). Therefore, educators can borrow from the evidence base that generally supports struggling adolescent writers (see review in Graham & Perin, 2007a, 2007b, 2007c). Using general interventions as a starting point will lead to the development and testing of viable discipline-specific science writing interventions designed to support struggling adolescents (Faggella-Luby et al., 2012).

Writing in the disciplines poses many challenges for struggling writers. Typically, struggling writers face difficulty with overall quantity of text production, organization of writing, content generation, motivation and perseverance, use of strategies especially for planning and revising, seeing the reader's perspective, and writing mechanics (Harris & Graham, 2009; Troia, 2006). With those challenges, students need additional supports beyond what is provided for other students in the class (e.g., explicit instruction of writing strategies [Harris & Graham, 2009; Graham, 2006] and other evidence-based writing adaptations [see Figure 1 quadrant 4; Graham, Olinghouse, & Harris, 2009; Graham & Perin, 2007a, 2007b, 2007c; Troia, 2006]).

The Current Study

This study examined teacher report of current instructional writing practice in secondary science classrooms in relationship to the Framework for Disciplinary Writing in Science to explore current implementation and potential areas for improvement. Survey methodology was used to collect teachers' reported practices to allow for a large-grain analysis of the writing practices currently taking place. The methodology revealed broad-stroke trends to identify areas for instructional reform and intervention (as in Cutler & Graham, 2008; Gilbert & Graham, 2010; Kiuhara et al., 2009) based on the recommendations of the Framework. As with previous survey studies of secondary writing instruction (Applebee & Langer, 2011; Gillespie et al., 2014; Graham et al., 2014; Kiuhara et al., 2009), this study serves as the first step in a programmatic line of research to inform the development of future observational (as in Pressley, Wharton-McDonald, Mistretta-Hampston, & Echevarria, 1998) and intervention studies—specifically in science writing. The research questions outline the focus of the study and begin to address the alignment among current practice and the Framework for Disciplinary Writing in Science.

Through teacher report, the following questions were explored based on the components of the Framework: (a) What are teachers' purposes for teaching writing in science class? (b) What types of writing do teachers assign in science class? (c) What evidence-based practices do teachers use to teach writing in science class? (d) What adaptations (if any) do teachers make for struggling writers? The last research question explored the fit of typical practice to the theoretical underpinnings of the disciplinary writing progressions: (e) How do these practices differ at the middle school level versus high school level?

Method

Participants

The sample was drawn from the population of currently practicing certified science teachers (Grades 6–12) in the United States (U.S.). Participants were selected using a random sampling procedure stratified by grade span (6–8 and 9–12). Three thousand certified science teachers from public elementary schools in the United States were randomly selected from a national database, Market Database Retrieval (MDR), of 97,760 secondary science teachers. A total of 3,000 teachers was determined to be more than appropriate for the sample size given the size of the population, a statistical confidence level of 95%, and a sampling error of ± 5, with an expected 15% response rate based on previous email surveys (Brindle, 2013) and some expected error within the list.

Procedures

The Tailored Design Method (Dillman, Smyth, & Christian, 2009) was used in this study to organize the design, implementation, and analysis of results, as it provides a step-by-step scientific procedure to reduce survey error and to optimize high quality response rates. With more than 3,600 professional publication citations, the Tailored Design Method employs multiple motivational features, enables alignment of all survey procedures, and helps to maximize response rate and quality of responses (Dillman et al., 2009). The Method also provided a resource to confirm the use of Web-based surveys as a viable single-mode delivery option for professional teachers who are expected to have regular access

[2] The positive effect of process writing on student outcomes is dependent on teachers' training in process writing approaches, such as the National Writing Project (Graham & Perin, 2007c). The majority of research on process writing has been conducted at the elementary level.

to e-mail and the Internet (Dillman et al., 2009; Hutchinson & Reinking, 2011; Kwak & Radler, 2002).

Using the Tailored Design Method as a guide, initial email requests to participate in the Qualtrics (http://qualtrics.com) survey were sent to the 2,003 science teachers whose names, e-mails, and positions could be confirmed from the original list of 3,000. Follow-up e-mail prompts were sent at the end of the first week and once a week for an additional four weeks (as recommended in Brindle, 2013 and Fraze, Hardin, Brashears, Haygood, & Smith, 2003) to those teachers who had not yet responded. After six weeks, the survey link was closed. As with print-based surveys, online surveys still need to contend with response rate as a primary issue of concern (Dillman et al., 2009). Participants were told that the first 300 participants to respond would receive a five-dollar Amazon gift card. Participants were assured that their responses remained anonymous throughout the entire survey process, including distribution of incentives.

Instrumentation

Teachers were surveyed via a 26-item questionnaire with sub items totaling 99 (instrument available as online supplemental resource). The questionnaire collected information about teachers' demographic background, preparation to teach writing, writing tasks assigned to students, use of evidence-based writing practices in science, and instructional adaptations used for particular students. All items on the questionnaire relate to elements of the Framework (see Figure 1). The instrument was developed after review of previous instruments that survey secondary teachers' reported writing practices (Applebee & Langer, 2011; Gillespie et al., 2014; Graham et al., 2014; Kiuhara et al., 2009).

Prior to dissemination of the questionnaire, a group of 20 experts were asked to participate in a construct validation of the instrument to ensure validity of the survey items for the purposes of this study. Seventy percent (n = 14) of the invited raters provided feedback on the items. The raters reflected the following areas of expertise: practicing science teachers (n = 2), practicing secondary special educator (n = 1), university faculty/researchers who are former science teachers (n = 3), university faculty/researchers who are former special educators (n = 2), university faculty/teacher educators with assessment expertise (n = 2), PhD candidates with literacy expertise (n = 3), a PhD candidate with measurement expertise, and a practicing psychometrician (n = 1).

The researchers employed quantitative and qualitative procedures (Netemeyer, Bearden, & Sharma, 2003) following McKenzie and colleagues' (1999) proposed steps for content validation to reword items for clarity and redundancy, determine representation of content within items, and to determine relevance of items. A content validity ratio (CVR) was calculated on each section of the instrument using McKenzie et al.'s (1999) recommended procedure. Items that did not meet the .62 cutpoint were removed (two items—writing brochures and posters in section II) or reworded (one item—theoretical papers in section II) based on feedback and strength of connection to the theoretical Framework.

An informal field test was conducted with two certified science teachers to determine the amount of time required to complete the questionnaire. Respondents were asked to focus on how long it took them to complete the instrument. The respondents reported completing the questionnaire in 12 and 19 min. They were also asked to report any difficulties they had with the survey platform. The respondents found two items that only allowed a single response where a multiple response choice was more appropriate. The questionnaire was modified to allow multiple response choices for those two items. No other problems with the questionnaire were reported.

Section I. Background information. The first 22 items of the questionnaire collected essential information on teachers' (a) personal and school characteristics, (b) preparation to teach writing (preservice and in-service), and (c) description of a typical class (from which they would base responses to the other sections of the questionnaire). Teachers were asked to report the average amount of time they spend teaching writing in science per week. They were also asked to report their purposes for teaching writing in science to answer RQ 1.

Section II. Writing assignments in science. Teachers were asked to rate how often they assign 21 different writing tasks in the science class they described in section I (e.g., lab report, argument, science notebook entry). This section responds to RQ 2. Teachers were asked to rate the frequency in which they assign each task using a seven-point Likert-type response format with the following options: 1 (*never*), 2 (*once a year*), 3 (*once a quarter*), 4 (*monthly*), 5 (*weekly*), 6 (*several times a week*), and 7 (*daily*).

Section III. Writing instruction in science. In 20 items, teachers were asked to rate the frequency with which they implement evidence-based instructional practices (e.g., writing strategies, inquiry activities, study of models) in response to RQ 3. As with section II, teachers rated the frequency of these practices with similar seven-point Likert-type response scale: 1 (*never*), 2 (*several times a year*), 3 (*monthly*), 4 (*several times a month*), 5 (*weekly*), 6 (*several times a week*), and 7 (*daily*).

Section IV. Adaptations for struggling writers. Participant responses to the 15 items in this section answered RQ 4. Teachers were asked to report the frequency in which they utilize the list of writing adaptations for struggling writers using the same seven-point Likert-type response scale used in section III.

Approach to Data Analysis

For sections II, III, and IV, respondents reported the frequency of writing practices on a seven-point Likert-type response scale, which was treated as interval and continuous for the purposes of analysis and interpretation (per de Winter & Dodou, 2012; McCoach, Gable, & Madura, 2013). Analysis served to identify areas of the theoretical Framework for Disciplinary Writing in Science that are and are not occurring in typical practice at the middle school and high school level.

Descriptive statistics are reported for the entire sample and for two subgroups: (a) middle school teachers and (b) high school teachers to provide a general overview of how often and in which capacity writing practices are being used in science class (RQ 2, 3, 4) and the differences between middle and high school teachers' use of the practices (RQ 5). In examining responses to teachers' identified purposes for teaching writing in science (RQ 1), a chi square analysis was used to determine whether there was an association between teachers' selected purpose and subgroup membership (middle or high school).

A series of factor analysis procedures were used to determine the number of factors to extract for items collected within section

II and within section III (RQ 2 and 3). Exploratory factor analyses (EFA) were run separately within each section because of the relatively large number of items and small number of respondents; as number of participants per item decreases, it becomes relatively unlikely that EFA will recover the correct factor structure. Costello and Osborne (2005) found that only 60% of samples will yield the correct factor structure when there are 10 participants per item, this reduces to 40% with five participants per item.

The Minimum Average Partial (MAP) procedure (Velicer, Eaton, & Fava, 2000), Parallel Analysis (PA) procedure (Hayton, Allen, & Scarpello, 2004), Eigenvalue greater than 1 test, and the scree plot test were conducted to explore the number of factors within each section. Principal axis factoring with oblique rotation through SPSS was conducted using the suggested factor solution. A priori, it was hypothesized that there are two factors within the writing assignment section (discipline-specific and general), and two factors within the evidence-based writing instruction section (writing to learn and learning to write). An iterative process was used to determine factor structure based on the above procedures, an examination of communalities, KMO, Bartlett's test of Sphericity, and multidimensional items (Pett, Lackey, & Sullivan, 2003). This process confirmed a two-factor solution on both sections II and III, however factor names were modified to adequately represent factor structure (see Results section). Correlations among the items on each factor were explored to confirm distinct categories, and diagonals above .50 on the anti-image correlation matrix were verified. Reliability analyses were conducted for each subscale to determine the Cronbach's alpha estimate.

Factor scores were calculated using the refined regression score method (DiStefano, Zhu, & Mindrila, 2009) to create a variable that reflected standardized regression weights similar to a z-score metric with a range of -3.0 to 3.0. To determine if differences existed, an analysis of variance (ANOVA) was conducted examining school level (middle or high school) difference on factor score-variables for each of the four scales for sections II and III. Then, follow up ANOVAs were conducted comparing item-level means for middle and high school, to further analyze differences. Significant differences at the scale level and at the item level are reported as effect sizes (d).

To confirm that treating the data as continuous, interval-level data with a normal distribution would yield the same results as treating it as if it violated those assumptions, the Mann–Whitney test was used and confirmed that similar results were yielded from nonparametric procedures. In all cases, except where noted, the parametric results were the same as nonparametric, and therefore only the parametric results are reported.

Results

Data were exported from Qualtrics survey management system and imported into an SPSS database. Only completed questionnaires were counted in the participant pool. In a few cases there were missing data on a given item, and therefore that participant was excluded from the analyses for that item. Although 2,003 teachers were sent the e-mail request to complete the study, the Qualtrics server reported that 1,969 teachers received the e-mail (the loss of 34 participants was determined to be attributable to error in the list or district firewalls). Of the 1,969 teachers who received an invitation to participate in the survey study, 287 teachers agreed to participate and completed the questionnaire. This constitutes a response rate of 14.5%, which met expectations for an online survey study using this methodology (Brindle, 2013). This response rate also met the qualifications Dillman and colleagues (2009) set forth, given the size of the population (N = 97,760), a statistical confidence level at 95%, and a slightly higher than recommended sampling error of ± 6 versus ± 5.

Respondents and nonrespondents were compared on three demographic variables (see Table 2). Chi-square analysis with a Bonferroni correction to adjust for Type I error [$\alpha = .05/3 = .017$] revealed no differences between respondents and nonrespondents on middle versus high school level [χ^2 (1, N = 2006) = 3.72, p = .05] or geographic region [χ^2 (4, N = 2006) = 1.29, p = .86]. There was a statistically significant difference on the variable of gender [χ^2 (1, N = 2006) = 8.37, p = .004], with women more likely to respond than men.

Demographic information on all respondents is reported to describe the school, teacher, student, and course characteristics of the participant pool (see Table 3). Teachers considered one class that represented how they typically include writing in science when reporting content and grade level. At the school level, 52% of teachers reported that their school was located in a suburban setting, with 27% indicating a rural setting, and 21% indicating an urban setting. Average school size across the group was 1,106 students. Forty-one percent of participants reported that their school is a Title 1 school. Participants have been teaching for an average of 13 years, and most participants have Master's degrees (73%). As a group, teachers reported limited preservice training in teaching writing in science. Only 9% of teachers took a preservice course that included writing instruction. However, many teachers reported that they received in-service professional development; the majority of teachers (56%) have attended professional development days on writing instruction in science.

Participants responded to the remainder of the instrument keeping in mind one particular science class that represented typical writing in science practice. Teachers described science classrooms in which the majority (57%) are targeted to learners at the "average" level (advanced placement, honors, basic, and heterogeneous were other response options). Of an average science class of 23 students (M = 23.40, SD = 12.61), teachers reported having about five students who receive special education services (M = 5.14, SD = 5.96) and three students who are English Language Learners (M = 3.07, SD = 9.36). The average length of the science block is 58 min (M = 57.68, SD = 17.34), and the portion of that block devoted to writing each week is on average only 15% or seven minutes per day (M = 6.74, SD = 7.14).

What Are the Purposes for Teaching Writing in Science?

Overall, teachers' purposes for teaching writing in science were distributed across the response options (see Table 4). The majority of teachers (59%) reported to include writing in science to accompany the scientific inquiry process (before, during, and after science investigations). The least frequently selected purpose was to meet state and national standards (16%). Only 4% of teachers reported not including writing in science. Pearson's chi-square test

Table 2
Characteristics of Respondents and Non-Respondents

Variable	Respondents *n*	Respondents %	Nonrespondents *n*	Nonrespondents %
School level				
Middle school	134	48%	933	54%
High school	146	52%	793	46%
Geographic region				
Northeast	82	29%	500	29%
Southeast	60	21%	385	22%
Midwest	63	23%	344	20%
West	42	15%	270	16%
Southwest	33	12%	227	13%
Gender				
Female	191	68%	1020	59%
Male	89	32%	706	41%

Note. Calculations computed on n (respondents) = 280 of N = 2,006 from original panel. Seven additional teachers participated through referral or snowball sampling.

revealed that the only purpose for which there was a significant difference between middle school teachers' responses and high school teachers' responses was "to review ideas and experiences," $\chi^2(1, N = 287) = 7.35, p = .007$. Despite a significant difference, the odds ratio (OR = .50) demonstrates a slight effect (Field, 2009).

What Types of Writing Do Teachers Assign in Science?

Table 5 displays how often teachers reported using the 21 writing assignment options in science class from most frequently assigned (highest mean) to the least frequently assigned (lowest mean). The majority of teachers reported to regularly assign (daily, several times per week, or weekly) note taking (92%), question/answers (75%), fill-in worksheets (69%), data recording sheets (60%), graphic organizers (60%), summaries (58%), and compare/contrasts (51%). A large portion of respondents reported *never* assigning correspondence (63%), digital texts (49%), scientific arguments (29%), research papers (28%), or theory papers (26%).

EFA revealed two underlying factors within the set of assignment items (see Approach to Data Analysis section). See Table 6 for factor loadings (pattern coefficients). A priori, it was hypothesized that these items would factor into two scales representing discipline-specific and general writing tasks (Faggella-Luby et al., 2012; Shanahan & Shanahan, 2008). Although a two-factor solution was supported, the groupings of items could not be explained as "discipline-specific" versus "general" writing tasks. The factor structure was further analyzed in order to name the scales accurately. The first factor was named 'External Audience Tasks.' This scale ($\alpha = .78$) includes the written work students do to communicate their scientific findings to an outside audience with the purpose of knowledge building. Students learn to write these types of tasks in science class to best communicate observations, findings, and theories to potential audiences beyond the teacher. The second factor was named 'Internal Audience Tasks' ($\alpha = .77$), and includes school-based tasks that use the act of writing to enhance memory and processing of science content. The audience for these tasks includes the student and sometimes the teacher. Correlation between the two factors was −.484.

Factor scores were computed for each respondent and the mean for middle and high school was compared to examine differences between school levels regarding the frequency in which writing tasks that focus on internal and external audiences are assigned. There was a statistically significant difference on both comparisons [External Audience $F(1, 257) = 8.33, p = .004, d = 0.36$ and Internal Audience $F(1, 257) = 14.46, p < .001, d = 0.47$], with middle school teachers scoring higher (standardized regression coefficients) on both scales, resulting in a difference with a medium effect (Hancock & Mueller, 2010). Further analysis was conducted on individual items to determine which items strongly influenced this difference.

Twenty-one separate nondirectional analyses were conducted with a one-way ANOVA using level of school (middle or high school) as the independent variable. Using the Bonferroni correction to control for Type I errors ($\alpha = .05/21$), the null hypothesis was rejected if the p value was less than .0023. Many of the items (9 of 21) did not meet the assumption of homogeneity of variance as evidenced by the Levene statistic. In those cases, the analysis was also run using Welch's F, which was designed for such cases. In all nine of those analyses, Welch's F and the ANOVA test

Table 3
Participant Characteristics

Variable	*n*	%
School locale		
Suburban	149	52
Rural	78	27
Urban	60	21
Education[a]		
Master's	210	73
Master's+	44	15
Content		
Biology	78	27
General Science	76	26
Other	40	14
Earth Science	38	13
Chemistry	54	12
Physics	20	7
Ethnicity		
White	261	92
Black	6	2
Other	9	3
Latino	6	2
Asian	3	1
Grade level		
6–8	138	49
9–10	80	28
11–12	46	16
Other	23	7
Teach Title I school	117	41
Took pre-service writing class	27	9
Attended writing professional development	159	56

Note. Some variables do not total n = 287 based on response for that particular item. Teachers selected a class that represented how they typically include writing in science, and reported grade level and content for that group of students.
[a] All teachers reported having either a Master's degree or a Master's degree plus additional credits beyond the Master's. This assumes they also have the requisite Bachelor's degree.

Table 4
Purposes for Teaching Writing in Science (n = 284)

Variable	n	%	χ^2	OR
To accompany the scientific inquiry process	168	59	.42	
Middle school	87	61		
High school	81	57		
To assess students	118	42	1.02	
Middle school	55	39		
High school	63	45		
To extend science learning	97	34	3.35	
Middle school	56	40		
High school	41	29		
To review ideas and experiences	96	34	*7.35	MS > HS: .50
Middle school	59	39		
High school	37	26		
To teach students to write scientific text	82	29	1.81	
Middle school	36	24		
High school	46	33		
To teach students scientific vocabulary	78	27	1.67	
Middle school	44	31		
High school	34	24		
To activate background knowledge	76	27	2.47	
Middle school	44	30		
High school	32	23		
To meet state/national standards	45	16	.02	
Middle school	23	16		
High school	22	16		
I do not include writing in science	11	4	.87	
Middle school	4	3		
High school	7	5		
Other	18	6	.98	
Middle school	6	5		
High school	12	8		

Note. 0 cells have expected frequencies <5. Odds ratio reported for statistical significance of chi-square comparisons. MS = middle school teachers; HS = high school teachers. Total percentages may not equal 100 because of rounding.
$^{*}p < .01$.

yielded identical results; therefore, only the ANOVA results are reported. Effect sizes (Cohen's d) are reported for each item with a statistically significant difference at the school level, taken by subtracting the high school mean from the middle school mean (see Table 5 for effect sizes).

For six of 21 writing tasks, there was a statistically significant difference between middle and high school teachers ($p < .0023$) with effect sizes between $d = 0.37$ and 0.71 reflecting a medium effect, approaching a large effect in a few instances (Hancock & Mueller, 2010). Middle school teachers reported assigning the following tasks more often than did high school teachers: hypotheses or theory papers [$F(1, 276) = 9.50$, $p = .0020$, $d = 0.37$], science notebooks [$F(1, 281) = 35.66$, $p < .001$, $d = 0.71$], summaries [$F(1, 278) = 11.41$, $p < .001$, $d = 0.40$], graphic organizers [$F(1, 277) = 19.74$, $p < .001$, $d = 0.54$], and written responses to science reading [$F(1, 282) = 22.03$, $p < .001$, $d = 0.55$]. High school teachers were more likely to assign note taking [$F(1, 281) = 11.54$, $p < .001$, $d = -0.40$].

What Evidence-Based Practices Do Teachers Use to Teach Writing in Science?

Table 8 illustrates how often teachers reported using the 20 evidence-based writing practices in science class from the most frequently utilized (highest mean) to the least frequently utilized (lowest mean) averaged across middle and high school teachers. The most commonly utilized instructional practices include teaching vocabulary, using writing as a tool for content learning, and providing class time for sustained silent writing. Yet, even these most frequently used practices are used daily or several times a week by less than 20% of teacher respondents. Other than those three practices, the remaining instructional practices were reported to be used by less than 7% of teachers either daily or several times a week. The least commonly assigned tasks included teaching strategies for revising and editing and explicitly teaching scientific genre. More than 30% of teachers report that they never explicitly teach strategies for planning writing or revising/editing writing, goals for what to include in writing, the elements/structure/style of scientific writing, the process approach to writing, and models of good writing.

EFA revealed two underlying factors within the set of instructional practices items (see Approach to Data Analysis section). Factor loadings (pattern coefficients) are presented in Table 7. The first factor represented 'Instructional Writing Practices' ($\alpha = .94$) and includes the approaches teachers use to teach writing in science. The second factor was named 'Writing Strategy Instruction' ($\alpha = .91$), and includes specific teaching approaches to teach

Table 5
Writing Tasks That Teachers Assign in Science Class

Item	Reported percentage of use of specific assignment											
	Never 1	1/Year 2	Qrterly 3	Mnthly 4	Weekly 5	Sev ×/ week 6	Daily 7	*Mdn*	*Mode*	*M*	*SD*	*d**
Note taking												
MS (*n* = 141)	0%	0%	2%	12%	38%	32%	16%	5	5	5.50	.946	
HS (*n* = 142)	1%	0%	1%	1%	28%	39%	30%	6	6	5.89	1.022	MS > HS: −.40
Question/Answers												
MS (*n* = 142)	2%	2%	2%	14%	33%	33%	15%	5	5	5.29	1.188	
HS (*n* = 141)	9%	1%	8%	11%	35%	19%	17%	5	5	4.88	1.671	
Fill-in worksheets												
MS (*n* = 140)	5%	1%	5%	21%	42%	21%	5%	5	5	4.78	1.292	
HS (*n* = 140)	11%	0%	7%	14%	34%	26%	9%	5	5	4.71	1.637	
Recording sheets												
MS (*n* = 142)	2%	0%	11%	22%	38%	24%	2%	5	5	4.74	1.147	
HS (*n* = 140)	5%	1%	8%	29%	39%	15%	4%	5	5	4.57	1.270	
Graphic organizers												
MS (*n* = 139)	2%	1%	3%	21%	41%	24%	8%	5	5	4.99	1.103	
HS (*n* = 140)	10%	4%	9%	28%	27%	17%	4%	4	4	4.26	1.567	MS > HS: .54
Summaries												
MS (*n* = 140)	4%	2%	5%	20%	45%	18%	6%	5	5	4.75	1.253	
HS (*n* = 140)	12%	3%	8%	30%	30%	14%	4%	4	4	4.18	1.561	MS > HS: .40
Compare/Contrasts												
MS (*n* = 141)	4%	1%	14%	29%	35%	15%	2%	5	5	4.44	1.197	
HS (*n* = 142)	12%	4%	6%	28%	35%	12%	4%	5	5	4.21	1.548	
Scientific Explanations												
MS (*n* = 138)	8%	2%	10%	29%	32%	16%	2%	4.5	5	4.28	1.414	
HS (*n* = 140)	11%	4%	13%	28%	34%	7%	3%	4	5	4.03	1.483	
Response to reading												
MS (*n* = 142)	4%	2%	11%	23%	36%	20%	5%	5	5	4.57	1.350	
HS (*n* = 142)	18%	6%	15%	26%	23%	10%	3%	4	4	3.73	1.668	MS > HS: .55
Science notebook												
MS (*n* = 142)	18%	2%	2%	13%	18%	16%	32%	5	7	4.75	2.223	
HS (*n* = 141)	47%	3%	4%	12%	14%	9%	11%	3	1	3.15	2.277	MS > HS: .71
Essay tests												
MS (*n* = 139)	8%	1%	13%	54%	19%	5%	1%	4	4	3.89	1.134	
HS (*n* = 143)	8%	2%	14%	42%	29%	4%	0%	4	4	3.94	1.203	
Lab reports												
MS (*n* = 143)	9%	7%	14%	41%	26%	3%	0%	4	4	3.77	1.237	
HS (*n* = 142)	8%	4%	23%	31%	30%	4%	1%	4	4	3.85	1.273	
Lists												
MS (*n* = 139)	19%	2%	9%	36%	23%	9%	2%	4	4	3.68	1.634	
HS (*n* = 140)	34%	1%	6%	29%	21%	7%	2%	4	1	3.33	1.844	
Description of models												
MS (*n* = 140)	19%	5%	17%	35%	17%	4%	2%	4	4	3.45	1.552	
HS (*n* = 140)	29%	6%	21%	23%	16%	4%	2%	3	1	3.09	1.662	
Procedural text												
MS (*n* = 141)	23%	5%	24%	27%	15%	4%	2%	3	4	3.26	1.534	
HS (*n* = 140)	24%	8%	24%	23%	11%	6%	3%	3	1	3.20	1.659	
PowerPoints												
MS (*n* = 140)	16%	17%	26%	27%	9%	2%	2%	3	3	3.10	1.400	
HS (*n* = 139)	21%	19%	19%	13%	16%	9%	4%	3	1	3.24	1.760	
Theory papers												
MS (*n* = 139)	20%	9%	22%	30%	19%	0%	1%	3	4	3.22	1.450	
HS (*n* = 139)	32%	13%	24%	18%	12%	1%	1%	3	1	2.68	1.470	MS > HS: .37
Scientific arguments												
MS (*n* = 141)	23%	12%	24%	26%	12%	2%	1%	3	4	2.99	1.417	
HS (*n* = 140)	35%	9%	19%	24%	9%	3%	1%	3	1	2.79	1.599	
Digital texts												
MS (*n* = 140)	46%	6%	12%	18%	9%	5%	4%	2	1	2.62	1.825	
HS (*n* = 139)	52%	10%	8%	12%	13%	5%	0%	1	1	2.40	1.709	
Research papers												
MS (*n* = 141)	21%	34%	29%	14%	1%	1%	1%	2	2	2.41	1.096	
HS (*n* = 141)	35%	31%	21%	9%	4%	1%	0%	2	1	2.17	1.146	
Correspondence												
MS (*n* = 140)	60%	11%	13%	11%	1%	3%	2%	1	1	1.96	1.434	
HS (*n* = 140)	66%	11%	6%	7%	7%	1%	1%	1	1	1.85	1.454	

Note. Percent totals may not equal 100. Effect size reported for statistically significant comparisons.
* $p > .0023$ (statistically significant at the .05 level after applying a Bonferroni adjustment to account for the large number of comparisons) when level was compared.

Table 6
Pattern Coefficients: Writing Tasks That Teachers Assign in Science Class

Item	Constructs: External audience tasks	Constructs: Internal audience tasks
Note taking	*	*
Question/Answers		**−.602**
Fill-in worksheets	—	—
Recording sheets	—	—
Graphic organizers		**−.548**
Summaries	.165	**−.560**
Compare/Contrasts		**−.665**
Scientific explanations	—	—
Response to reading	—	—
Science notebook	.390	−.177
Essay tests	—	—
Lab reports	**.429**	.111
Lists		**−.626**
Description of models	**.556**	−.110
Procedural text	**.395**	−.149
PowerPoints	—	—
Theory papers	**.713**	
Scientific arguments	**.612**	−.114
Digital texts	**.430**	−.203
Research papers	**.601**	
Correspondence	**.492**	

Note. Pattern coefficients loading at or above .395 are bolded to illustrate relationship with factor. Any item that loaded on both factors above .275 or on neither factor at or above .390 was omitted from EFA (—). Items with communalities below .20 were removed prior to EFA (*). Coefficients below .10 are suppressed (blank). See discussion of negative factor loadings in text (Discussion).

students strategies for effective writing. Fifteen items strongly loaded on the first factor and five items very strongly loaded on the second factor. Correlations between the two factors was −.732.

As with the previous section, factor scores were computed for each respondent to support subgroup analyses for each factor. To determine whether differences existed between middle and high school teachers, an ANOVA was conducted examining school level difference on factor score-variables for both 'Instructional Writing Practices' and 'Writing Strategy Instruction.' A statistically significant difference was observed between middle and high school teachers on both factors—'Instructional Writing Practices' [$F(1, 251) = 15.35$, $p < .001$, $d = 0.49$] and 'Writing Strategy Instruction' [$F(1, 251) = 15.88$, $p < .001$, $d = 0.50$]. When the strength of that difference was examined, it resulted in a medium effect size of $d = 0.49$ for 'Instructional Writing Practices' and $d = 0.50$ for 'Writing Strategy Instruction.' Individual analyses were conducted to determine key aspects of the difference.

Thus, 20 separate nondirectional ANOVA analyses were conducted using level of school (middle or high school) as the independent variable. Using the Bonferroni correction to control for Type I errors ($\alpha = .05/20$), the null hypothesis was rejected if the p value was less than .0025. For 13 of 20 instructional writing practices, there was a statistically significant difference between middle school and high school teachers ($p < .0025$) with medium effect sizes between $d = 0.37$ to 0.57. Middle school teachers reported using all of the following evidence-based practices more often than high school teachers: teach strategies for planning [$F(1, 276) = 19.87$, $p < .001$, $d = 0.53$], teach strategies to organize writing [$F(1, 274) = 14.39$, $p < .001$, $d = 0.46$], and teach strategies to summarize [$F(1, 276) = 13.71$, $p < .001$, $d = 0.44$]; teach students to establish goals [$F(1, 277) = 11.25$, $p < .001$, $d = 0.40$]; have students collaborate when writing [$F(1, 276) = 10.68$, $p < .001$, $d = 0.39$]; have students engage in prewriting activities [$F(1, 276) = 22.69$, $p < .001$, $d = 0.57$]; have students emulate good models [$F(1, 275) = 10.62$, $p < .001$, $d = 0.39$]; have students use writing as a tool for learning [$F(1, 273) = 13.47$, $p < .001$, $d = 0.45$]; have students self-regulate writing toward goals [$F(1, 272) = 9.98$, $p = .0020$, $d = 0.39$]; use explicit instruction [$F(1, 274) = 20.64$, $p < .001$, $d = 0.55$]; have students conduct inquiry projects [$F(1, 275) = 9.50$, $p = .0020$, $d = 0.37$]; provide class time for silent sustained writing [$F(1, 274) = 10.46$, $p < .001$, $d = 0.39$]; and teach students genre-specific vocabulary [$F(1, 272) = 13.36$, $p < .001$, $d = 0.44$].

What Adaptations Do Teachers Make for Struggling Writers in Science?

Table 9 presents the frequency with which teachers report using the 15 evidence-based adaptations identified in the theoretical framework. Science teachers reported using the following writing supports most frequently (weekly or more frequently): providing a distraction-free environment (33%), additional time to write (24%), additional prompting (23%), and a word bank or glossary (22%). Science teachers use the following practices least often (at least weekly): specialized technology (8%), teach grammar or mechanics (10%), and conference with students (5%). Thirty percent or more of science teachers reported that they *never* use

Table 7
Pattern Coefficients: Evidence-Based Instructional Writing Practices Teachers Use in Science Class

Item	Constructs: Writing instruction	Constructs: Strategy instruction
Teach students vocabulary	**.749**	.194
Use writing as tool to learn	**.602**	−.114
Provide class time for writing	**.717**	
Provide teacher/peer feedback	**.744**	
Use technology for writing	**.591**	
Collaborate when writing	—	—
Conduct inquiry projects	**.660**	
Engage prewriting activities	**.533**	−.248
Self-regulate towards goals	**.592**	−.198
Teach strategies: summarize		**−.706**
Engage in research process	**.794**	
Teach strategies: organize		**−.869**
Use explicit instruction	**.586**	−.236
Use a process approach	**.652**	−.109
Emulate models of good writing	**.584**	−.223
Establish goals for writing	.117	**−.721**
Teach students to reference	**.739**	
Teach strategies: Planning		**−.778**
Teach strategies: Edit/Revise		**−.804**
Explicitly teach scientific text	**.566**	−.151

Note. Pattern coefficients loading above .399 are bolded to illustrate relationship with factor. Any item that loaded on both factors above .299 or on neither factor above .399 was omitted from EFA (—). Coefficients below .10 are suppressed (blank). See Discussion of negative factor loadings in text.

Table 8
Evidence-Based Instructional Writing Practices Teachers Use in Science Class

	Reported percentage of use of specific practice											
Item	Never 1	Sev ×/ Year 2	Monthly 3	Sev ×/ Month 4	Weekly 5	Sev ×/ Week 6	Daily 7	*Mdn*	*Mode*	*M*	*SD*	*d**
Teach students vocabulary												
MS (*n* = 137)	15%	21%	6%	13%	21%	16%	8%	4	2	3.81	1.957	
HS (*n* = 137)	28%	23%	11%	17%	9%	6%	6%	2	1	2.97	1.843	MS > HS: .44
Use writing as tool to learn												
MS (*n* = 138)	15%	21%	15%	18%	13%	12%	6%	3	2	3.49	1.813	
HS (*n* = 137)	26%	29%	15%	18%	7%	1%	4%	2	2	2.73	1.593	MS > HS: .45
Provide class time for writing												
MS (*n* = 139)	14%	23%	19%	14%	15%	9%	5%	3	2	3.38	1.750	
HS (*n* = 137)	22%	32%	16%	18%	7%	2%	3%	2	2	2.74	1.510	MS > HS: .39
Provide teacher/peer feedback												
MS (*n* = 138)	13%	28%	18%	20%	11%	9%	2%	3	2	3.18	1.548	
HS (*n* = 136)	14%	42%	15%	13%	10%	4%	2%	2	2	2.84	1.482	
Use technology for writing												
MS (*n* = 140)	15%	32%	11%	21%	13%	5%	3%	3	2	3.14	1.615	
HS (*n* = 140)	24%	31%	16%	17%	6%	2%	4%	2	2	2.69	1.540	
Collaborate when writing												
MS (*n* = 138)	15%	28%	13%	22%	13%	6%	3%	3	2	3.16	1.581	
HS (*n* = 140)	25%	32%	19%	14%	6%	3%	1%	2	2	2.57	1.415	MS > HS: .39
Conduct inquiry projects												
MS (*n* = 140)	12%	29%	19%	16%	18%	4%	2%	3	2	3.10	1.528	
HS (*n* = 137)	25%	31%	19%	18%	4%	2%	1%	2	2	2.56	1.371	MS > HS: .37
Engage pre-writing activities												
MS (*n* = 138)	13%	33%	13%	17%	13%	9%	2%	3	2	3.14	1.576	
HS (*n* = 140)	35%	33%	12%	11%	6%	2%	1%	2	1	2.29	1.376	MS > HS: .57
Self-regulate towards goals (*n* = 274)												
MS (*n* = 137)	20%	30%	13%	15%	11%	8%	3%	2	2	2.99	1.671	
HS (*n* = 137)	31%	31%	15%	16%	4%	1%	1%	2	1	2.40	1.374	MS > HS: .39
Teach strategies: summarize												
MS (*n* = 140)	18%	30%	16%	19%	12%	2%	2%	3	2	2.90	1.485	
HS (*n* = 138)	28%	41%	13%	10%	6%	1%	1%	2	2	2.29	1.251	MS > HS: .44
Engage in research process												
MS (*n* = 137)	13%	41%	17%	17%	6%	4%	2%	2	2	2.76	1.364	
HS (*n* = 139)	22%	45%	13%	11%	6%	3%	0%	2	2	2.43	1.286	
Teach strategies: organize												
MS (*n* = 139)	13%	40%	14%	15%	13%	2%	3%	2	2	2.89	1.488	
HS (*n* = 137)	26%	47%	12%	9%	3%	2%	1%	2	2	2.27	1.222	MS > HS: .46
Use explicit instruction												
MS (*n* = 139)	20%	33%	16%	15%	9%	5%	2%	2	2	2.83	1.546	
HS (*n* = 137)	41%	31%	14%	10%	4%	1%	0%	2	1	2.07	1.186	MS > HS: .55
Use a process approach												
MS (*n* = 138)	20%	39%	14%	14%	6%	4%	2%	2	2	2.64	1.474	
HS (*n* = 140)	39%	31%	13%	9%	5%	2%	0%	2	1	2.16	1.293	
Emulate models of good wrtg												
MS (*n* = 139)	24%	31%	17%	12%	10%	3%	2%	2	2	2.68	1.509	
HS (*n* = 138)	38%	36%	9%	9%	6%	1%	0%	2	1	2.13	1.266	MS > HS: .39
Establish goals for writing												
MS (*n* = 139)	27%	27%	16%	14%	11%	2%	2%	2	2	2.66	1.535	
HS (*n* = 140)	41%	35%	10%	9%	3%	1%	2%	2	1	2.09	1.328	MS > HS: .40
Teach students to reference												
MS (*n* = 138)	24%	41%	13%	13%	6%	4%	1%	2	2	2.48	1.357	
HS (*n* = 137)	29%	42%	13%	10%	2%	3%	1%	2	2	2.26	1.273	
Teach strategies: planning												
MS (*n* = 139)	21%	38%	10%	16%	10%	2%	2%	2	2	2.69	1.488	
HS (*n* = 139)	40%	42%	8%	5%	2%	2%	1%	2	2	1.97	1.185	MS > HS: .53
Teach strategies: edit/revise												
MS (*n* = 138)	28%	41%	13%	8%	6%	3%	1%	2	2	2.32	1.351	
HS (*n* = 140)	41%	41%	9%	6%	1%	2%	0%	2	1	1.94	1.107	
Explicitly teach scientific text												
MS (*n* = 138)	33%	38%	10%	9%	3%	4%	2%	2	1	2.28	1.470	
HS (*n* = 139)	42%	36%	11%	6%	4%	1%	1%	2	1	1.97	1.179	

Note. Percent totals may not equal 100. Effect size reported for statistically significant comparisons.
* $p > .0025$ (statistically significant at the .05 level after applying a Bonferroni adjustment to account for the large number of comparisons) when school level was compared.

Table 9
Adaptations Teachers Use in Science Class to Support Struggling Writers

Item	Reported percentage of use of specific practice: Never 1	Sev ×/ Year 2	Monthly 3	Sev ×/ Month 4	Weekly 5	Sev ×/ Week 6	Daily 7	*Mdn*	*Mode*	*M*	*SD*	*d**
Additional time to write												
MS (n = 137)	13%	29%	13%	10%	16%	12%	7%	3	2	3.50	1.863	
HS (n = 140)	30%	28%	14%	15%	5%	3%	5%	2	1	2.66	1.670	MS > HS: .48
Small group instruction												
MS (n = 138)	28%	25%	10%	15%	11%	6%	4%	2	1	2.86	1.760	
HS (n = 140)	43%	24%	12%	11%	6%	1%	2%	2	1	2.26	1.506	MS > HS: .37
A distraction-free environment												
MS (n = 137)	12%	19%	17%	10%	15%	15%	12%	4	2	3.86	1.918	
HS (n = 139)	27%	26%	10%	11%	9%	6%	11%	2	1	3.09	2.025	MS > HS: .39
A word bank/glossary												
MS (n = 136)	17%	17%	19%	14%	12%	10%	11%	3	1	3.54	1.970	
HS (n = 140)	43%	21%	14%	9%	6%	3%	4%	2	1	2.39	1.682	MS > HS: .63
A rubric/checklist												
MS (n = 137)	13%	31%	16%	19%	10%	6%	5%	3	2	3.14	1.610	
HS (n = 140)	21%	34%	18%	16%	7%	1%	2%	2	2	2.68	1.416	
Additional explicit instruction												
MS (n = 139)	13%	30%	13%	18%	18%	5%	3%	3	2	3.19	1.619	
HS (n = 139)	34%	28%	17%	14%	4%	1%	1%	2	1	2.37	1.384	MS > HS: .55
Strategies for organizing text												
MS (n = 139)	18%	36%	12%	16%	13%	4%	2%	2	2	2.85	1.503	
HS (n = 140)	25%	46%	14%	10%	4%	0%	2%	2	2	2.30	1.251	MS > HS: .40
Additional prompting												
MS (n = 138)	10%	31%	10%	19%	17%	5%	8%	3	2	3.41	1.729	
HS (n = 140)	16%	32%	22%	12%	11%	3%	4%	3	2	2.91	1.538	
Graphic organizers												
MS (n = 139)	13%	25%	17%	18%	16%	8%	4%	3	2	3.36	1.642	
HS (n = 139)	29%	27%	17%	13%	9%	3%	2%	2	1	2.63	1.543	MS > HS: .46
Models of good writing												
MS (n = 138)	10%	33%	14%	18%	14%	6%	5%	3	2	3.24	1.610	
HS (n = 140)	23%	36%	19%	13%	4%	2%	2%	2	2	2.54	1.385	MS > HS: .47
Tasks broken into components												
MS (n = 137)	18%	29%	10%	19%	14%	8%	2%	3	2	3.06	1.662	
HS (n = 138)	28%	28%	18%	17%	7%	1%	1%	2	1	2.55	1.404	
Peer help for struggling writer												
MS (n = 139)	38%	29%	14%	8%	7%	2%	2%	2	1	2.35	1.508	
HS (n = 141)	50%	26%	9%	9%	4%	0%	2%	1	1	1.99	1.352	
Teacher conferences												
MS (n = 139)	26%	40%	14%	13%	5%	1%	2%	2	2	2.40	1.300	
HS (n = 141)	35%	41%	11%	9%	1%	1%	1%	2	2	2.09	1.195	
Grammar/mechanics instr.												
MS (n = 140)	37%	27%	9%	13%	8%	2%	4%	2	1	2.46	1.651	
HS (n = 141)	52%	23%	14%	4%	4%	2%	1%	1	1	1.94	1.308	
Specialized technology												
MS (n = 136)	70%	11%	5%	2%	7%	3%	2%	1	1	1.81	1.518	
HS (n = 141)	74%	11%	6%	4%	3%	1%	1%	1	1	1.58	1.237	

Note. Percent totals may not equal 100. Effect size reported for statistically significant comparisons.
* $p > .003$ (statistically significant at the .05 level after applying a Bonferroni adjustment to account for the large number of comparisons) when school level was compared.

specialized technology, teach grammar or mechanics, provide additional explicit instruction, give students a rubric or checklist, provide students with a word bank or glossary, or offer small group instruction.

Results were compared for middle and high school teachers using 15 separate ANOVAs with school level (middle or high) as the independent variable. To adjust for Type I error, the Bonferroni correction was used (α = .05/15), and the null hypothesis was rejected if the p value was less than .0033. With eight of the 15 practices, there was a statistically significant difference between middle and high school teachers at the $p < .0033$ level. The resulting effect sizes reflected medium effects ranging from d = 0.37 to 0.63. Middle school teachers reported using all of the following adaptations more often than high school teachers: teach students ways of organizing text [$F(1, 277) = 11.00, p < .001, d = 0.40$], have students use graphic organizer when writing [$F(1, 276) = 14.74$, $p < .001$, $d = 0.46$], provide models of good writing [$F(1, 276) = 14.95$, $p < .001$, $d = 0.47$], provide additional explicit/direct instruction [$F(1, 276) = 20.97, p < .001, d = 0.55$], provide small group instruction [$F(1, 276) = 9.28$, p =

.0030, $d = 0.37$], provide additional time to write [$F(1, 275) = 15.87$, $p < .001$, $d = 0.48$], provide a distraction-free environment [$F(1, 274) = 10.45$, $p < .001$, $d = 0.39$], and provide students with a word bank or glossary [$F(1, 274) = 27.66$, $p < .001$, $d = 0.63$].

Discussion

Purposes for Teaching Writing in Science

In science writing instruction informed by the recommendations of the Framework, teachers acknowledge how writing aligns to the overarching goals of science education, and make purposeful decisions to include writing in science class. Study results confirmed that participating teachers purposefully include writing in science. Participants indicated a range of diverse purposes for including writing in science matching the overarching goals of science education. The purpose with the greatest research base as highlighted in the Framework matched what the majority of teachers reported as a reason to include writing in science: 59% of teachers reported that they include writing in science to accompany the inquiry process. However, the other two most relevant purposes highlighted in the Framework were selected by less than a third of the teachers: only 29% of teachers selected to teach students how to write scientific text, and only 27% selected to teach students scientific vocabulary. In moving ahead, the gap between typical practice and the recommendations of the Framework needs to be more closely considered, especially since the Framework aligns with the writing required of students in the new science standards (Achieve, 2013). Teaching students how to write scientific text will help them to be more effective in communicating scientific understanding in writing—toward the ultimate purpose of knowledge building in science class (Moje, 2015; NRC, 2012; Scardamalia & Bereiter, 2006).

Tasks Teachers Assign in Science

Recommendations based on the Framework include teachers assigning analytical tasks to promote deep learning of science concepts and to communicate scientific understanding to authentic audiences. Yet, the most frequently assigned tasks reported by survey participants are restricted writing tasks that require very little composition, and are thus less likely to promote learning. Teachers most often reported assigning note taking, questions and answers, fill-in worksheets, data recording sheets, and graphic organizers. Although this result matches results from previous national surveys (Applebee & Langer, 2011; Gillespie et al., 2014; Kiuhara et al., 2009), it does not align to the full recommendations of the Framework and its supporting research base.

Teachers do report assigning a range of tasks in science class. All but five of the 21 task options are assigned more frequently than once a quarter on average. This result is promising. However, according to the research and theory behind the Framework, not all tasks equally promote learning in science. Assigning tasks that allow students to analyze and synthesize information and utilize discipline-specific genres are more likely to promote deep learning and help students contribute to a scientific learning community, which is the goal of science education (NRC, 2012; Scardamalia & Bereiter, 2006). So although teachers are including a variety of writing tasks in science, they are not regularly including tasks that are most relevant to promoting learning or communicating scientific findings to an authentic audience. For example, theory papers and scientific arguments, two of the five least frequently assigned tasks, have the strongest body of evidence to support their use to promote learning and analysis and to communicate scientific findings to contribute to the knowledge of a scientific learning community.

The Framework builds from a polytheoretical foundation that emphasizes a progression of writing instruction that culminates in discipline-specific, knowledge building tasks assigned frequently at the high school level (Scardamalia & Bereiter, 1991, 2006, 2010; Shanahan & Shanahan, 2008, 2012). The results of the exploratory factor analyses (EFA) for Section II of the instrument (writing assignments) began to confirm the distinction of tasks as articulated by the Framework. Although discipline-specific tasks did not emerge on its own factor, most of the tasks on the ‘External Audience’ tasks scale reflected analytic writing tasks that are specific to the discipline of science (lab reports, theory papers, procedural text, scientific argument, description of scientific models).

The results of the EFA were also used to generate factor scores that allowed for further analysis. A comparison of respondents' factor scores on ‘External Audience’ tasks and ‘Internal Audience’ tasks revealed an interesting, yet concerning finding. In part illustrating the negative correlation ($r = -.484$), teachers who frequently use ‘internal audience’ tasks such as list making and questions and answers are less likely to frequently assign ‘external audience’ tasks such as scientific arguments and theory papers. Most teachers are infrequently assigning both categories of writing tasks, yet, teachers who tend to include more ‘external audience’ tasks use ‘internal audience’ tasks less frequently, and visa versa. The large-grain results of this study are unable to provide reasons for why this relationship exists. Future research is needed to confirm this preliminary relationship, as well as to further investigate possible reasons and counterfactuals. One hypothesis is that there is an underlying relationship between the writing assignments teachers select to use in science class and teachers' theoretical orientations about writing. A study such as the one conducted by Graham and colleagues (2002) that examined the relationship between teachers' practices and their beliefs and theories about writing and instruction, could be used to further investigate this relationship by using an adapted version of the *Writing Orientations Scale* along with the current instrument.

Evidence-Based Practices Teachers Use to Teach Writing in Science

Approaches aligned to the Framework for Disciplinary Writing in Science can help teachers to improve writing and content outcomes for all learners through the use of evidence-based writing practices. Evidence-based writing practices specifically for the science discipline have not yet emerged in the research, but there are numerous instructional writing practices that have been shown to improve adolescent outcomes across disciplines (Graham & Perin, 2007a, 2007b, 2007c). The practices should not be considered as add-ons, but can fit within an inquiry-based pedagogy focused on the authentic practices of scientists (NRC, 2012) that teachers use to teach science. Without these evidence-based practices, teachers may be assigning writing in science, but not spend-

ing ample time to teach students how to be successful with those writing tasks. It is the use and application of the evidence-based practices that is critical for helping to improve struggling learners' writing and content outcomes.

Survey results from this study confirmed what previous survey studies have indicated (Applebee & Langer, 2011; Kiuhara et al., 2009). The recommended evidence-based practices, as captured in the Framework, are not consistently making their way into secondary classrooms. According to the results of the current study, each practice is only taught at a frequency between once per year and once per quarter. There are four practices that are done quarterly on average: teach students vocabulary, use writing as a tool for learning, provide class time to write, and provide teacher or peer feedback. Whether done yearly or quarterly, there is a sizable gap between the expectations of the Framework and what is actually happening in science classrooms.

Exploratory factor analysis indicated that strategy instruction factored out separately from the other evidence-based writing practices. This finding supports substantial research over the past three decades (Graham, 2006) that has elevated strategy instruction as one of the most effective evidence-based practices for writing instruction. Strategy instruction has been shown to produce very large effect sizes on writing quality across research findings (d = .82 as reported in Graham and Perin's meta-analysis, 2007a.)—yet, it is reported to be used at least weekly by only 11% of survey respondents. Moreover, strategy instruction is never used by a third of the respondents in the sample. A comparison of respondents' factor scores on the 'Instructional Writing Practices' factor and the 'Writing Strategy Instruction' factor revealed an interesting pattern. Although most respondents are doing little of either type of instruction, those who more frequently use other instructional writing practices are less likely to use strategy instruction. This is another finding that will need attention in future research. In the literature, strategy instruction is a high-yield evidence-based writing practice. One interpretation that makes sense from these findings (along with the negative correlation of $r = -.732$ between the factors) is that the 'Writing Strategy Instruction' factor actually represents a lack of strategy instruction, which is reinforced by the negative factor loadings (pattern coefficients). With this interpretation, teachers who use other instructional writing practices more frequently would be more likely to include strategy instruction.

The overall lack of frequently used instructional writing practices and strategy instruction in science class will be critical to address when further studying or implementing the Framework. A closer look at the barriers to regularly implementing these practices will be necessary in moving ahead.

Adaptations Teachers Make for Struggling Writers in Science

The literature base supporting disciplinary literacy has not yet considered the implications for struggling adolescent learners (Faggella-Luby et al., 2012). Yet, students with disabilities and other needs often require additional supports to meet writing expectations (see a general review on writing adaptations in Graham, Harris, Fink-Chorzempa, & MacArthur, 2003). Results from this survey study confirm what other survey studies have found (Graham et al., 2003; Kiuhara et al., 2009)—that teachers make adaptations for struggling writers infrequently.

Teachers reported about one third of their typical class is comprised of students who receive special education services or are English Language Learners. However, most adaptations are being made by the majority of teachers on an annual or quarterly basis. Fewer than 3% of teachers surveyed use graphic organizers, rubrics, small group instruction, and/or conferences to scaffold their instruction on a daily basis. Only 10% of teachers use additional explicit instruction daily. On average, nearly a third of the teachers surveyed never use these highlighted adaptations.

Middle School Versus High School Teachers' Practices

On the whole, middle school teachers more regularly assign certain writing tasks, implement evidence-based practices for teaching writing, and use instructional adaptations to support writing in science class. On 26 of 56 comparisons, middle school teachers more regularly implemented writing tasks and instructional and adaptation practices. All but two aspects of Framework that middle school teachers report to implement more frequently than high school teachers (hypothesis/theory papers and science notebooks) reflect general, versus disciplinary, writing practices, which align with the theoretical perspective described earlier in this article. Articulated within Disciplinary Writing Progressions, middle school teachers' primary focus is to teach students how to become proficient writers within the content of science in order to use writing in a 'knowledge transforming' approach (Bereiter & Scardamalia, 1987), so that writing becomes a vehicle for science learning (internal audience). High school teachers, on the other hand, are expected to teach students to use advanced and science-specific (e.g., disciplinary) writing for the purpose of 'knowledge building' (Bereiter & Scardamalia, 2010), or communicating scientific understanding in the form of theories that explain and predict phenomena (Chuy et al., 2010) to a broader (external audience) scientific community. According to results of this study, middle school teachers are more aligned to the goals of the Framework, specifically the Disciplinary Writing Progressions. One possible explanation is that high school teachers do not yet see how writing and science are synergistic (Pearson et al., 2010), or how writing is linked to the overarching goals of science education and the nature of science. The Next Generation Science Standards may help teachers make the connection between writing and science through the Science and Engineering Practices, which include the writing practices of explanation and argument, and also allow for writing to be paired with other aspects of scientific investigation.

Although middle school teachers from this study report to be teaching writing aligned with the Disciplinary Writing Progressions, not all middle school students are entering high school ready to use writing for 'knowledge building' (NCES, 2012). Many high school students have tremendous writing needs that require the general writing supports described in the Framework. In fact, the gap between struggling and on-grade-level adolescent learners is greater in high school than it is in middle school (Deshler & Hock, 2006). To help all students meet the increased writing demands of college and career, high school teachers can regularly use evidence-based writing practices and instructional adaptations within a comprehensive program of regular assessment and appropriate interventions to support the needs of struggling writers.

Recommendations Based on the Framework for Disciplinary Writing in Science

The discrepancies highlighted above provide a lens through which to focus the priorities for improving writing instruction in science class. Therefore, the following changes to practice are recommended.

Consider diverse and distinct purposes for writing in science that align to the overarching goals of science education. Based on these survey results, there does not seem to be a clear consensus on the purpose for including writing in science except to accompany the inquiry process. As reported in previous survey studies, science teachers are not confident to teach writing (Applebee & Langer, 2011; Gillespie et al., 2014; Kiuhara et al., 2009). Because of that lack of confidence, they may be more willing to assign writing as assessments or to accompany the scientific inquiry process. Yet, for students to develop as science writers, they need explicit instruction in the forms, techniques, and elements of writing so they can write to communicate deep understanding through the use scientific genres and language/vocabulary. The Next Generation Science Standards will also help science teachers begin to understand that writing is an important component in the work of practicing scientists (Achieve, 2013), and should therefore be highlighted in science education.

Assign writing tasks in science to promote deep learning and to communicate scientific understanding to authentic audiences. The science teachers from this study more frequently assign restricted writing tasks for the 'internal audience' of the writer. The most frequently assigned tasks do not include extended student composition, analysis, or synthesis. This emphasis misses the purpose of science education, which is to see writing as a key aspect of the scientific inquiry process, and to use writing to communicate findings to a broader audience through knowledge building. More frequently assigned writing tasks, such as fill-in worksheets, may support learning of scientific concepts, but not deep understanding of the nature of science (Lederman, 2007). The value of scientific argument may not be fully emphasized in science classrooms, as it is one of the least frequently assigned writing tasks. This study confirms the work of Chuy and colleagues (2010) in the claim that high school teachers have not fully realized the value of writing to enhance disciplinary knowledge building in science class. Practice aligned to the Framework prioritizes analytical writing tasks in science specifically focusing on asking students to write lengthier compositions for audiences and purposes authentic to the discipline of science. This is especially important at the high school level, where students can use writing to build and extend knowledge within a scientific learning community.

Use general evidence-based practices to teach writing during science class. There are evidence-based instructional practices that teachers can confidently employ to improve students' writing and learning outcomes. As reported in this survey study and others (Applebee & Langer, 2011; Gillespie et al., 2014; Kiuhara et al., 2009), these practices are not being used regularly enough (and less frequently at the high school vs. the middle school level) to help students learn to regulate their own writing behavior. Cognitive strategy instruction is a proven way to help students grow as writers. Increasing the presence and frequency of writing strategy instruction within inquiry pedagogy may be an efficient and effective way to better support writers within science class. High school teachers have a great opportunity and responsibility to include more evidence-based writing practices in science class. These practices seem to offer a promising option for managing increased college and career demands on students who are struggling to meet minimal requirements.

Use evidence-based adaptations to support struggling writers in science. Similar to previous survey studies (Kiuhara et al., 2009), teachers in this study reported infrequently using adaptations to support struggling writers, despite the prevalence of students who may struggle with writing in their classrooms. Overall, a range of adaptations are not being used regularly for struggling writers, especially at the high school level. Evidence-based adaptations can provide struggling writers with the support they need. This will require teachers to increase the presence and frequency of adaptations to support struggling writers in science class, especially at the high school level. Many of the issues with decreased motivation at the high school level may be solved by better meeting students' needs through scaffolded instruction, while also promoting authentic tasks for real-world audiences and purposes.

Limitations

Although the study was designed to allow for generalization of results to teachers nationwide, several limitations remain. First and perhaps most importantly, the research team's estimates of the frequency with which writing instruction occurs in secondary science classrooms may overestimate the extent to which writing is taught because of social desirability bias. Teachers may have elevated their frequency estimates. Because the survey focuses on writing instruction in science, teachers will likely infer that the researcher team values this and therefore tend to inflate their estimates (Baumann, Hoffman, Duffy-Hester, & Moon Ro, 2000). This concern is somewhat offset by the fact that Web based instruments are less likely to be affected by social desirability bias than face-to-face or telephone interviews (Dillman et al., 2009). Other sources of response bias might also occur, including memory issues, inattentiveness to questionnaire items, or fatigue (Brown, 2004).

Despite efforts to ensure that results were generalizable, including randomly sampling participants, using incentives to increase response rates, sending reminder e-mails about the survey and promoting the survey through reliable sources (i.e., affiliating it with a university and organization that promotes science education), there is still the possibility that results cannot be generalized across the entire population of science teachers in the United States. First, teachers were only asked to report on one class, which may or may not have reflected typical practice across their full teaching load. Second, our sample underrepresents chemistry and physics teachers. Results might differ by scientific field and therefore skew findings.

Finally, results from the Exploratory Factor Analysis need to be considered as truly exploratory. This is the first study to use the Writing Assignments in Science scale in its entirety. Further research is needed to confirm the factor structure of the instrument.

Moving Ahead

It is recommended that findings from this study be shared with National Science Teachers Association (NSTA), with the goal of

creating a position paper on the role of writing in science instruction within a standards-based climate of accountability. Furthermore, it is recommended that examples of how writing instruction can be embedded within science inquiry pedagogy be created and made available to teachers through publication and professional development. Findings from this study can also be shared with science educators for inclusion in secondary science methods courses as well as literacy educators for inclusion in content methods coursework for preservice teachers.

The overall lack of use of evidence-based writing practices in science class indicates a need to strengthen core classroom writing instruction in science. It is recommended that follow-up intervention studies using a design experiment methodology be explored to meet this need (Bulgren & Ellis, 2012; Burkhardt & Schoenfeld, 2003; Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Gersten, Baker, & Lloyd, 2000). The Framework for Disciplinary Writing in Science provides the research community with a structure upon which discipline-specific writing interventions can be developed, so that a clear line of research can grow into a robust evidence base. In addition to studies focused on intervention design, another line of research that could emerge from this study is a set of observation studies on teachers' writing practices in science. Observation data would provide depth and nuance to the findings of this current study. A sampling methodology for an observation study could include identifying teachers who score high on the current instrument for follow-up observations and interviews.

Final Words

A multitheoretical Framework for Disciplinary Writing in Science explains how writing can be used to deepen students' learning in science in a way that promotes knowledge building and students' written contribution to scientific learning communities. The Framework provides the first comprehensive model of its kind to articulate how writing can be used in science to support student learning outcomes for a range of learners including struggling adolescent learners. This study also contributes to an overall lack of literacy research on discipline-specific practices at the high school level. Most importantly, the instruction and adaptation practices brought forth in the Framework consider all students, including struggling adolescent learners and students with disabilities. The previous literature on writing in science did not include these populations. This study is the first comprehensive national survey to examine teachers' behaviors related to teaching writing specifically in science.

References

Achieve, Inc. (2013). *Next generation science standards*. Washington, DC: Achieve, Inc. Retrieved from http://www.nextgenscience.org/

Applebee, A. N. (1981). *Writing in the secondary school: English and the content areas* (Research Monograph No. 21). Urbana, IL: National Council of Teachers of English.

Applebee, A. N. (1984). *Contexts for learning to write: Studies of secondary school instruction*. Norwood, NJ: Ablex.

Applebee, A. N., & Langer, J. A. (2011). A snapshot of writing instruction in middle schools and high schools. *English Journal, 100,* 14–27.

Archer, A., & Hughes, C. (2011). *Explicit instruction: Effective and efficient teaching*. New York, NY: Guilford Press.

Bangert-Drowns, R., Hurley, M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74,* 29–58. http://dx.doi.org/10.3102/00346543074001029

Baumann, J., Hoffman, J., Duffy-Hester, A., & Moon Ro, J. (2000). The first R yesterday and today: U.S. elementary reading instruction practices reported by teachers and administrators. *Reading Research Quarterly, 35,* 338–377. http://dx.doi.org/10.1598/RRQ.35.3.2

Baxter, G., Bass, K., & Glaser, R. (2001). Notebook writing in three fifth-grade science classrooms. *The Elementary School Journal, 102,* 123–140. http://dx.doi.org/10.1086/499696

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. London, UK: Erlbaum.

Bereiter, C., & Scardamalia, M. (2010). Can children really create knowledge? *Canadian Journal of Learning and Technology, 36*. Retrieved from http://www.cjlt.ca/index.php/cjlt/article/view/585

Brindle, M. (2013). *Examining relationships among teachers' preparation, efficacy, and writing Practices* (Unpublished dissertation). Department of Special Education, Vanderbilt University, Nashville, TN. Retrieved from http://etd.library.vanderbilt.edu/available/etd-06092013-102827/unrestricted/BrindleDissertation.pdf

Britton, J. (1970). *Language and learning*. New York, NY: Penguin.

Brown, G. T. (2004). Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. *Psychological Reports, 94,* 1015–1024. http://dx.doi.org/10.2466/pr0.94.3.1015-1024

Bulgren, J. A., & Ellis, J. D. (2012). Argumentation and evaluation intervention in science classes: Teaching and learning with Toulmin. In M. S. Kline (Ed.), *Perspectives on scientific argumentation: Theory, practice, and research* (pp. 135–154). New York, NY: Springer http://dx.doi.org/10.1007/978-94-007-2470-9_8

Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher, 32,* 3–14. http://dx.doi.org/10.3102/0013189X032009003

Cervetti, G., Barber, J., Dorph, R., Pearson, P. D., & Goldschmidt, P. G. (2012). The impact of an integrated approach to science and literacy in elementary school classrooms. *Journal of Research in Science Teaching, 49,* 631–658. http://dx.doi.org/10.1002/tea.21015

Chuy, M., Scardamalia, M., Bereiter, C., Prinsen, F., Resendes, M., Messina, R., . . . Chow, A. (2010). Understanding the nature of science and scientific progress: A theory-building approach. *Canadian Journal of Learning and Technology, 36,* 1–21.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *American Educational Research Association, 32,* 9–13.

Cook, B. G., & Cook, S. C. (2013). Unraveling evidence-based practices in special education. *The Journal of Special Education, 47,* 71–82. http://dx.doi.org/10.1177/0022466911420877

Cook, B. G., Tankersley, M., Cook, L., & Landrum, T. J. (2008). Evidence-based practices in special education: Some practical considerations. *Intervention in School and Clinic, 44,* 69–75. http://dx.doi.org/10.1177/1053451208321452

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10,* 1–9. Retrieved from http://pareonline.net/pdf/v10n7.pdf

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology, 100,* 907–919. http://dx.doi.org/10.1037/a0012656

Deshler, D., & Hock, M. (2006). Adolescent literacy: Where we are—Where we need to go. *LD Online*. Retrieved from http://www.ldonline.org/article/12288/

de Winter, J., & Dodou, D. (2012). Five-point Likert items: *t* test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation, 15*. Retrieved from http://pareonline.net/getvn.asp?v=15&n=11

Dillman, D., Smyth, J., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design model* (3rd ed.). Hoboken, NJ: Wiley

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, & Evaluation, 14,* 1–7. Retrieved from http://pareonline.net/pdf/v14n20.pdf

Drew. (2013). *Literature review of writing practices in science classrooms, grades 4–12* (Unpublished comprehensive exam). Storrs, CT: Department of Educational Psychology, University of Connecticut.

Emig, J. (1977). Writing as a mode of learning. *College Composition and Communication, 28,* 122–128. http://dx.doi.org/10.2307/356095

Faggella-Luby, M., Graner, P. S., Deshler, D., & Drew, S. V. (2012). Building a house on sand: Why disciplinary literacy is not sufficient to replace general strategies for adolescent learners who struggle. *Topics in Language Disorders, 32,* 69–84. http://dx.doi.org/10.1097/TLD.0b013e318245618e

Fang, Z., & Coatoam, S. (2013). Disciplinary literacy: What you want to know about it. *Journal of Adolescent & Adult Literacy, 56,* 627–632. http://dx.doi.org/10.1002/JAAL.190

Fang, Z., & Schleppegrell, M. J. (2010). Disciplinary literacies across content areas: Supporting secondary reading through functional language analysis. *Journal of Adolescent & Adult Literacy, 53,* 587–597. http://dx.doi.org/10.1598/JAAL.53.7.6

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage

Fraze, S., Hardin, K., Brashears, M., Haygood, J., & Smith, M. (2003). The effects of delivery mode upon survey response rate and perceived attitudes of Texas agri-science teachers. *Journal of Agricultural Education, 44,* 27–37. http://dx.doi.org/10.5032/jae.2003.02027

Gaskins, I., Guthrie, J., Satlow, E., Ostertag, L. S., Byrne, J., & Connor, B. (1994). Integrating instruction of science, reading, and writing: Goals, teacher development, and assessment. *Journal of Research in Science Teaching, 31,* 1039–1056. http://dx.doi.org/10.1002/tea.3660310914

Gersten, R., Baker, S., & Lloyd, J. W. (2000). Designing high-quality research in special education: Group experimental design. *The Journal of Special Education, 34,* 2–18. http://dx.doi.org/10.1177/002246690003400101

Gilbert, J., & Graham, S. (2010). Teaching writing to elementary students in grades 4–6: A national survey. *The Elementary School Journal, 110,* 494–518. http://dx.doi.org/10.1086/651193

Gillespie, A., Graham, S., Kiuhara, S., & Hebert, M. (2014). High school teachers' use of writing to support students' learning: A national survey. *Reading and Writing, 27,* 1043–1072. http://dx.doi.org/10.1007/s11145-013-9494-8

Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187–207). New York, NY: Guilford Press.

Graham, S., Capizzi, A., Harris, K., Hebert, M., & Morphy, P. (2014). Teaching writing to middle school students: A national survey. *Reading and Writing, 27,* 1015–1042. http://dx.doi.org/10.1007/s11145-013-9495-7

Graham, S., Harris, K., Fink-Chorzempa, B., & MacArthur, C. (2003). Primary grade teachers' instructional adaptations for struggling writers: A national survey. *Journal of Educational Psychology, 95,* 279–292. http://dx.doi.org/10.1037/0022-0663.95.2.279

Graham, S., Harris, K., MacArthur, C., & Fink, B. (2002). Primary grade teachers' theoretical orientations concerning writing instruction: Construct validation and a nationwide survey. *Contemporary Educational Psychology, 27,* 147–166. http://dx.doi.org/10.1006/ceps.2001.1085

Graham, S., Olinghouse, N., & Harris, K. (2009). Teaching composition to students with learning disabilities: Scientifically supported recommendations. In G. Troia (Ed.), *Instruction and assessment for struggling writers: Evidence-based practices* (pp. 165–186). New York, NY: Guilford Press.

Graham, S., & Perin, D. (2007a). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99,* 445–476. http://dx.doi.org/10.1037/0022-0663.99.3.445

Graham, S., & Perin, D. (2007b). What we know, what we still need to know: Teaching adolescents to write. *Scientific Studies of Reading, 11,* 313–335. http://dx.doi.org/10.1080/10888430701530664

Graham, S., & Perin, D. (2007c). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools: A report to Carnegie Corporation of New York.* Washington, DC: Alliance for Excellent Education.

Hancock, G. R., & Mueller, R. O. (Eds.). (2010). *The reviewer's guide to quantitative methods in the social sciences.* New York, NY: Routledge.

Hand, B. M., & Prain, V. (2002). Teachers implementing writing-to-learn strategies in junior secondary science: A case study. *Science Education, 86,* 737–755. http://dx.doi.org/10.1002/sce.10016

Harris, K., & Graham, S. (2009). Self-regulated strategy development in writing: Premises, evolution, and the future. *British Journal of Educational Psychology Monograph Series, 2,* 113–135. http://dx.doi.org/10.1348/978185409X422542

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7,* 191–205. http://dx.doi.org/10.1177/1094428104263675

Hutchinson, A., & Reinking, D. (2011). Teachers' perceptions of integrating information and communication technologies into literacy instruction: A national survey in the United States. *Reading Research Quarterly, 46,* 312–333.

Keys, C., Hand, B., Prain, V., & Collins, S. (1999). Using the science writing heuristic as a tool for learning from laboratory investigations in secondary science. *Journal of Research in Science Teaching, 36,* 1065–1084. http://dx.doi.org/10.1002/(SICI)1098-2736(199912)36:10<1065::AID-TEA2>3.0.CO;2-I

Kiuhara, S., Graham, S., & Hawken, L. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology, 101,* 136–160. http://dx.doi.org/10.1037/a0013097

Klein, P., & Kirkpatrick, L. (2010). A framework for content area writing: Mediators and moderators. *Journal of Writing Research, 2,* 1–46. http://dx.doi.org/10.17239/jowr-2010.02.01.1

Klein, P., & Rose, M. (2010). Teaching argument and explanation to prepare junior students for writing to learn. *Reading Research Quarterly, 45,* 433–461. http://dx.doi.org/10.1598/RRQ.45.4.4

Kwak, N., & Radler, B. (2002). A comparison between mail and web surveys: Response pattern, respondent profile, and data quality. *Journal of Official Statistics, 18,* 257–273.

Langer, J., & Applebee, A. (2007). *How writing shapes thinking: A study of teaching and learning.* WAC Clearinghouse Landmark Publications in Writing Studies. Retrieved from http://wac.colostate.edu/books/langer_applebee/. Originally Published in Print, 1987, by National Council of Teachers of English, Urbana, IL.

Lederman, N. G. (2007). Nature of science: Past, present, and future. In S. Abell & N. Lederman (Eds.), *Handbook of research on science education* (pp. 831–879). New York, NY: Routledge.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Evidence based on the internal structure of the instrument: Factor analysis. In D. McCoach, R. Gable, & J. Madura (Eds.), *Instrument development in the affective domain: School and corporate applications* (3rd ed., pp. 109–161). New York, NY: Springer Science and Business Media. http://dx.doi.org/10.1007/978-1-4614-7135-6_4

McKenzie, J. F., Wood, M. L., Kotecki, J. E., Clark, J. K., & Brey, R. A. (1999). Establishing content validity: Using qualitative and quantitative

steps. *American Journal of Health Behavior, 23,* 311–318. http://dx.doi.org/10.5993/AJHB.23.4.9

McNeill, K. L. (2011). Students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. *Journal of Research in Science Teaching, 48,* 793–823. http://dx.doi.org/10.1002/tea.20430

McNeill, K. L., & Krajcik, J. (2009). Synergy between teacher practices and curricular scaffolds to support students in using domain specific and domain general knowledge in writing arguments to explain phenomena. *Journal of the Learning Sciences, 18,* 416–460. http://dx.doi.org/10.1080/10508400903013488

Moje, E. B. (2015). Doing and teaching disciplinary literacy with adolescent learners: A social and cultural enterprise. *Harvard Educational Review, 85,* 254–278. http://dx.doi.org/10.17763/0017-8055.85.2.254

National Center for Educational Statistics. (2016). *The condition of education 2016* (NCES 2016–144). Washington, DC: U.S. Department of Education, Institute of Educational Sciences.

National Center for Education Statistics. (2012). *The nation's report card: Writing 2011* (NCES 2012–470). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Commission on Writing in America's Schools and Colleges. (2003, April). *The neglected R: The need for a writing revolution.* Retrieved from http://www.collegeboard.com/prod_downloads/writingcom/neglectedr.pdf

National Research Council. (1996). *National science education standards.* Washington, DC: The National Academy Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Committee on a Conceptual Framework for New K–12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications.* Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985772

Newell, G. (2006). Writing to learn: How alternate theories of school writing account for student performance. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 235–247). New York, NY: Guilford Press.

Pearson, P. D., Moje, E., & Greenleaf, C. (2010). Literacy and science: Each in the service of the other. *Science, 328,* 459–463. http://dx.doi.org/10.1126/science.1182595

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research.* Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412984898

Prain, V., & Hand, B. (1996). Writing for learning in science: A model for use within classrooms. *Australian Science Teachers Journal, 42,* 23–27.

Pressley, M., Wharton-McDonald, R., Mistretta-Hampston, J., & Echevarria, M. (1998). Literacy instruction in 10 fourth and fifth grade classrooms in upstate New York. *Scientific Studies of Reading, 2,* 159–194. http://dx.doi.org/10.1207/s1532799xssr0202_4

Santangelo, T., & Olinghouse, N. (2009). Effective writing instruction for students who have writing difficulties. *Focus on Exceptional Children, 42,* 1–20.

Scardamalia, M., & Bereiter, C. (1991). Higher levels of agency for children in knowledge building: A challenge for the design of new knowledge media. *Journal of the Learning Sciences, 1,* 37–68. http://dx.doi.org/10.1207/s15327809jls0101_3

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 97–118). New York, NY: Cambridge University Press.

Scardamalia, M., & Bereiter, C. (2010). A brief history of knowledge building. *Canadian Journal of Learning and Technology, 36,* 1–16.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review, 78,* 40–59. http://dx.doi.org/10.17763/haer.78.1.v62444321p602101

Shanahan, T., & Shanahan, C. (2012). What is disciplinary literacy and why does it matter? *Topics in Language Disorders, 32,* 7–18. http://dx.doi.org/10.1097/TLD.0b013e318244557a

Starr, M., & Krajcik, J. (2013). Developing and using models to align with NGSS. *Science Scope, 37,* 31–35. http://dx.doi.org/10.2505/4/ss13_037_01_31

Troia, G. (2006). Writing instruction for students with learning disabilities. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 324–336). New York, NY: Guilford Press.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helms (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41–71). Norwell, MA: Springer. http://dx.doi.org/10.1007/978-1-4615-4397-8_3

Received September 25, 2015
Revision received November 15, 2016
Accepted December 13, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 7, 956–976

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000189

Measuring Arithmetic: A Psychometric Approach to Understanding Formatting Effects and Domain Specificity

Katherine T. Rhodes
Ohio State University

Lee Branum-Martin and Julie A. Washington
Georgia State University

Lynn S. Fuchs
Vanderbilt University

Using multitrait, multimethod data, and confirmatory factor analysis, the current study examined the effects of arithmetic item formatting and the possibility that across formats, abilities other than arithmetic may contribute to children's answers. Measurement hypotheses were guided by several leading theories of arithmetic cognition. With a sample of 1,314 third grade students (age M = 103.24 months, SD = 5.41 months), Abstract Code Theory, Encoding Complex Theory, Triple Code Theory, and the Exact versus Approximate Calculations Hypothesis were evaluated, using 11 measures of arithmetic with symbolic problem formats (e.g., Arabic numeral and language-based formats) and various problem demands (e.g., requiring both exact and approximate calculations). In general, results provided support for both Triple Code Theory and Encoding Complex Theory. As predicted by Triple Code Theory, arithmetic outcomes with language formatting, Arabic numeral formatting, and estimation demands (across formats) were related but distinct from one another. As predicted by Encoding Complex Theory, executive attention was a direct predictor of all arithmetic outcomes. Language was no longer a direct predictor of arithmetic outcomes when executive attention was accounted for in the model; however, a strong and enduring relationship between language and executive attention suggested that language may play a facilitative role in reasoning during numeric processing. These findings have important implications for assessing arithmetic in educational settings and suggest that in addition to arithmetic-focused interventions, interventions targeting executive attention, language, and/or the interplay between them (i.e., internal speech during problem-solving) may be a promising avenues of mathematical problem-solving intervention.

Educational Impact and Implications Statement

Symbolic formats (e.g., Arabic numerals, spoken language, written language) are usually used for teaching and testing arithmetic ability in formal educational settings; however, research has suggested that different symbolic formats may lead to different sorts of arithmetic problem-solving. Using a large sample of elementary school-aged children, this study explored the possibility that the manner in which problems are conveyed during testing may be an important factor for understanding arithmetic cognition and achievement. Findings suggested that (1) different types of symbolically-formatted arithmetic problems measure different constellations of skills, and (2) symbolic formats may not be appropriate for measuring an ability that is purely mathematical. Executive attention was a significant and direct predictor of arithmetic performance across problem formats. Language ability was not a direct predictor of arithmetic performance, but rather appeared to facilitate executive attention, helping students maintain attention and coordinate problem-solving procedures. These findings have important implications for the selection and interpretation of arithmetic assessments that are commonly used in educational settings. For example, students experiencing difficulty with word problems are likely struggling with understanding concepts like selecting appropriate strategies and executing the procedural steps of the strategies they select, and to a lesser extent may also be struggling with concepts like interpreting number words. Findings also suggest that targeting executive attention and/or language-facilitated executive attention (i.e., internal speech) during mathematical problem-solving may be promising avenues of intervention.

This article was published Online First March 23, 2017.

Katherine T. Rhodes, Department of Psychology, Ohio State University; Lee Branum-Martin, Department of Psychology, Georgia State University; Julie A. Washington, Department of Educational Psychology, Special Education, and Communication Disorders, Georgia State University; Lynn S. Fuchs, Department of Special Education, Vanderbilt University.

This article was prepared, in part, for fulfillment of the requirements of a doctoral dissertation for KTR. This research was supported by Award R24HD075454, R24HD075443, and R01HD059179 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health. We thank the participating students, teachers, and schools who made this research possible.

Correspondence concerning this article should be addressed to Katherine T. Rhodes, Department of Psychology, Ohio State University, Columbus, OH 43210. E-mail: rhodes.390@osu.edu

Keywords: arithmetic cognition, functional numeracy, mathematics achievement testing, common method variance, symbolic formatting

Supplemental materials: http://dx.doi.org/10.1037/edu0000189.supp

Arithmetic mastery is essential for successful daily living and is foundational for advanced-level participation in STEM disciplines (American Association on Intellectual and Development Disabilities, 2010; Coalition, S. T. E. M., 2000). Despite decades of mathematics education reform, children in the United States continue to struggle with math achievement, and this is true of both basic arithmetic skills and more advanced problem solving (National Center for Education Statistics, 2013; Woodward, 2004). This study explored the possibility that problem formatting, the manner in which problems are conveyed during testing, may be an important factor for understanding arithmetic cognition and achievement.

Format-Based Concerns for Word Problems

Formatting of assessment stimuli is an important consideration for the measurement of arithmetic ability (Ansari, 2007; Campbell, 1994; Dehaene, Piazza, Pinel, & Cohen, 2003; Lourenco, Bonny, Fernandez, & Rao, 2012; McCloskey, 1992; Piazza, Pinel, Le Bihan, & Dehaene, 2007). Symbolic formats (e.g., Arabic numerals, spoken language, and written language) are usually used for teaching and testing arithmetic ability in formal educational settings; however, research has suggested that different symbolic formats may lead to different sorts of mental representation and processing of numerical information (Ansari, 2007; Campbell, 1994; Dehaene et al., 2003).

In the realm of educational testing, linguistic formats serve an important purpose for testing arithmetic ability. Language formats are often used to convey everyday "word problems" in a variety of testing scenarios. For example, both the National Assessment of Educational Progress (NAEP) and the Program for International Student Assessment (PISA) use "word problems" to assess students' understandings of real-world mathematics (Kelly et al., 2013; National Center for Education Statistics, 2013). Word problems are generally thought to go beyond basic arithmetic knowledge, testing students' abilities to apply their conceptual knowledge and strategic competence to problem-solving situations they encounter in and outside of the classroom (Greer, 1997; National Academy of Sciences, 2001; Verschaffel, De Corte, & Lasure, 1994). Ideally, word problems require students to both decide upon strategies for problem-solving and apply their arithmetic and procedural knowledge to execute those strategies. Thus, linguistic formats represent a valuable means of surmising how students will perform arithmetic in their daily lives.

Despite their importance for assessing arithmetic ability, word problems have been the subject of measurement controversy, perhaps because they became common as indicators of mathematical ability on popular standardized testing instruments. They have been criticized for the extent to which they fail to encourage students to apply common-sense to mathematical problem-solving (Baranes, Perry, & Stigler, 1989; Verschaffel, Greer, & de Corte, 2000), conversely for the extent to which they may penalize students with less world or situational knowledge (Chipman, Marshall, & Scott, 1991; Davis-Dorsey, Ross, & Morrison, 1991; Stern & Lehrndorfer, 1992), and perhaps most notably for penalizing students with lower reading ability (Ballew & Cunningham, 1982; Helwig, Rozek-tedesco, & Tindal, 2002; Helwig, Rozek-tedesco, Tindal, Heath, & Almond, 1999; Muth, 1984).

Efforts in test reform have largely acknowledged these concerns about content, design, and administration of linguistically formatted items (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985, 1999, 2014; National Research Council Committee on Appropriate Test Use, 1999). However, more recently, a number of researchers have suggested that linguistic formats may unintentionally tap executive abilities (e.g., working memory) and language ability, particularly when examinees are unfamiliar with the language system utilized in test formatting (Abedi & Lord, 2001; Martiniello, 2009; Rhodes, Branum-Martin, Morris, Romski, & Sevcik, 2015; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Terry, Hendrick, Evangelou, & Smith, 2010). This line of research is focused on the idea that the very formatting of word problems may be a source of inherent testing bias.

To be clear, the newer line of research focused on issues of testing bias for linguistically formatted mathematics problems is not the only research tradition to implicate domain general abilities as important to arithmetic cognition. Several educational and developmental researchers have implicated executive abilities and language in arithmetic performance (e.g., Bull, Espy, & Wiebe, 2008; Bull & Scerif, 2001; Cummins, Kintsch, Reusser, & Weimer, 1988; Hecht, Torgesen, Wagner, & Rashotte, 2001; LeFevre et al., 2013; Mazzocco & Kover, 2007; Passolunghi, Vercelloni, & Schadee, 2007; Zheng, Swanson, & Marcoulides, 2011). Furthermore, the idea that both conceptual knowledge and procedural/strategic ability contribute to successful performance on word problems is not new (see e.g., Nesher, 1986; Riley, Greeno, & Heller, 1983; Siegler, 1991; Siegler & Shrager, 1984). However, the question of domain general testing bias is distinct from the question of domain general contributions. This difference is subtle but important. From the valid measurement/testing bias perspective, the issue is not whether domain general abilities are important contributors to arithmetic development. Rather, the issue is whether commonly used measures of arithmetic ability are actually also measures of domain general abilities and should be interpreted as such (in which case it would not be surprising that measures of "arithmetic ability" correlate with or can be linearly regressed upon measures of domain general abilities).

Detecting Format-Based "Bias" in Arithmetic Problems

Even with careful design of problem content, formatting may pose a hidden threat to the validity of a measure (Messick,

1989, 1996). Though "bias" is a term that usually means "unfair" or "discriminatory" in popular speech, it generally refers to the underlying issue of construct validity in psychometric contexts (Crocker & Algina, 2008; Hambleton, Swaminathan, & Rogers, 1991; Reynolds & Suzuki, 2012). A test item is biased when two examinees with the same level of ability would not have the same probabilities of correctly answering (often called "differential item functioning" or DIF; see e.g., Borsboom et al., 2002; Hambleton et al., 1991). The unequal probability of correct answers is always because the item is unintentionally measuring some dimension other than the one intended by test developers (i.e., it is not unidimensional). When the bias is an artifact of the way test items are formatted, it can more specifically be referred to as common method variance (CMV; Cote & Buckley, 1987, 1988; Messick, 1989, 1996; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). In this case, the linguistic formatting of word problems may lead to the unintentional measurement of language and executive abilities in addition to the intentional measurement of arithmetic abilities, and the interpretation of test scores as indicators of math ability regardless of formatting would be invalid.

Detecting common method variance requires a specific measurement methodology, which ideally is guided by a strong theoretical foundation. Potential confounding dimensions of language and executive abilities must be measured along with arithmetic, in a variety of formats (i.e., a multitrait, multimethod methodology). Then measures of arithmetic ability and potential confounding dimensions must be included in statistical models of responses, which evaluate not only mean structures but also variance structures. This can be accomplished with statistical models capable of allowing for the possibility that multiple abilities may predict behaviors (see e.g., Cote & Buckley, 1987; Eid, Lischetzke, & Nussbeck, 2006; Marsh, Beard, & Bailey, 2002; Maul, 2013). These statistical models fall under the broad umbrella of factor analysis, and in the case where they are theoretically guided and specified by a priori hypotheses about construct measurement, they may be more specifically referred to as "confirmatory factor analysis."

Forming Measurement Hypotheses With Leading Theories of Arithmetic Cognition

Four leading theories of arithmetic cognition, Abstract Code Theory, Encoding Complex Theory, Triple Code Theory, and Exact versus Approximate Calculations, provide the theoretical basis for forming confirmatory factor models of arithmetic cognition in the current study. Each of these theories attempts to explain (a) how we encode numerical information and represent numerical information mentally, (b) how we retrieve math facts from memory, process the information, or operate upon numerical representations to achieve solutions to problems, (c) how we recode our mental, numerical representations of solutions into output and report our answers, and (d) which cognitive domains are involved in these activities and how they may interact with one another, if at all. In general, these four facets define the process of "arithmetic" for theories of arithmetic cognition, and each of these facets of arithmetic are areas in which theories of arithmetic cognition may diverge from each other, sometimes irreconcilably.

One consequence of this theoretical divergence has been that there is no consensus assessment of numeric processing. Cognitive research on arithmetic has considered any number of calculation demands represented in a variety of formats as potentially valid measures of arithmetic. Across various arithmetic tasks, dominant theories of arithmetic cognition offer different accounts of how people do arithmetic, what mental processes are involved in arithmetic, and why people exhibit individual differences in arithmetic ability. The following section considers each theory with regard to its specifications for the process and measurement of arithmetic, with special attention as to how it attempts to explain language-formatting effects and the roles that language and executive abilities may play in arithmetic performance.

Abstract Code Theory

Abstract Code Theory stipulates that a single, *abstract code* is used to mentally represent all numeric information, regardless of input format (e.g., Arabic numerals; McCloskey, 1992; McCloskey, Caramazza, & Basili, 1985). Because this *abstract*, semantic *code* is the object of numeric processing, formatting exerts no effect on numeric processing (McCloskey, 1992). Differences in reaction time (RT) seen with language-formatted arithmetic stimuli can be attributed to increased encoding time necessary for mental representation of language input (McCloskey, Macaruso, & Whetstone, 1992). The extent to which a *language* domain may be involved in aiding numeric *comprehension* or *production* is unclear and not specified by the theory, but rather addressed as an area for future investigation (McCloskey, 1992). Similarly, the theory does not specify the extent to which some *executive system of control* (regulation, attention, inhibition, and working memory) is responsible for coordinating numeric *comprehension*, *processing*, and *production*. Rather, Abstract Code Theory tends to allow for a specialized numeric *processing* module to facilitate the execution of arithmetic operations. McCloskey (1992) notes that the roles of general processing abilities (e.g., working memory) are issues for future investigation.

Encoding Complex Theory

Encoding Complex Theory stipulates that the presentation of numerical stimuli activates an associative network of format-specific numerical "codes" or mental representations (Campbell, 1994; Campbell & Clark, 1988; Clark & Campbell, 1991). Mental representations of number can be verbal (e.g., articulatory, orthographic) or nonverbal (e.g., visual, motor, and analog magnitude; Campbell, 1994; Campbell & Clark, 1988; Clark & Campbell, 1991). The mental representations or "codes" are associatively connected within a complex network, called the *encoding complex*, and as such, they are assumed to stimulate each other in complex patterns of activation without the use of a common, abstract code (Campbell & Clark, 1988; Clark & Campbell, 1991). Successful numeric processing (number comprehension, calculation, comparison, and parity judgment) requires enhancing relevant association patterns and inhibiting interfering association patterns within the *encoding complex* network, and this is particularly true for calculation activities (Campbell & Clark, 1988; Clark & Campbell, 1991).

Encoding Complex Theory does not specify a specific *quantitative* domain as responsible for numeric processing. Instead, Campbell and Clark (1988; Clark & Campbell, 1991) have implicated a number of domain general cognitive capacities in resolving the complex network of associations activated during numeric processing. These domains include *executive systems of control* (inhibition, problem-solving, attention, working memory, and specifically, Baddeley & Hitch's, 1974 model of working memory), the *motor* domain, the *language* domain, and the *visuospatial* domain. Though *executive systems of control* are implicated across problem-solving activities and *language* ability is implicated in language formatted problems, the roles of *motor* and *visuospatial* abilities in predicting outcomes across various formats and relating to other cognitive domains during problem-solving is unclear.

Triple Code Theory

According to Triple Code Theory, stimulus format affects encoding and mental representation of number. The format in which number stimuli are presented will determine the type of mental representation encoded for them. Arabic numeral input is represented by the *visual Arabic number form*; language-based numeral input is represented by the *verbal word frame*; sets of objects are represented by the *analogical magnitude representation* (Dehaene, 1992; Dehaene & Cohen, 1995; Dehaene et al., 2003). Although each of these factors is allowed to communicate directly with one another via transcoding, problem demands influence the way in which numerical processing is conducted. Under Triple Code Theory, format-based differences in arithmetic performance are thus attributed to issues of efficiency in the transcoding process (Campbell & Epp, 2005).

The cognitive factors responsible for encoding and mentally representing numeric information are not the only cognitive domains involved in Triple Code Theory's arithmetic. The *language* domain supports the recognition of spoken and written number input, the production of spoken and written number output, and the retrieval of number facts (e.g., two plus two equals four) from memory (Dehaene, 1992; Dehaene & Cohen, 1995). The role of *executive systems* in coordinating the functions of arithmetic is unclear in Triple Code Theory. Although the three factors for the mental representation of number are assumed to cooperate with one another and with the *language* domain in carrying out numeric processing, the extent to which their cooperation is self-directed as opposed to organized by a super ordinate system of attention, inhibition, working memory, and regulation is not specified by the theory.

Exact Versus Approximate Calculations, an Extension of Triple Code Theory

Unlike the other theories of arithmetic cognition, exact versus approximate calculations theory pertains specifically to the numeric processing task of calculations. It is an extension of Triple Code Theory, supporting the idea that distinct neural networks contribute to (a) approximate calculation tasks involving semantic representations of quantity, comparison, and estimation versus (b) exact calculation tasks involving the retrieval of rote, verbal, numerical facts about quantity to compute exact arithmetic solutions (Dehaene et al., 1999; Stanescu-Cosson et al., 2000). The *analogical magnitude representation* domain is hypothesized to be supported by the neural network for approximate calculations, and the *verbal word frame* domain is hypothesized to be supported by the neural network for exact calculations. These domains appear to be integrated, and they may both be recruited for difficult, exact calculation problems involving large quantities (Stanescu-Cosson et al., 2000).

Other assumptions of Triple Code Theory, including the possible cognitive domains involved in numeric processing are generally not addressed in the empirical literature supporting exact versus approximate calculations. The focus of this empirically generated theory is specifying the roles of the *analogical magnitude representation* domain and the *verbal word frame* domain on approximate and exact calculation activities. The *visual Arabic number form* domain is largely absent from this specification of Triple Code Theory; however, spatial attention networks, possibly representing some of the predictive power of the *visual Arabic number form* domain and possibly representing some form of *executive control* for attention, may contribute to coordinating both types of task.

Summary: Comparing and Contrasting Theories of Arithmetic Cognition

Although Abstract Code, Encoding Complex, Triple Code, and Exact versus Approximate Calculations Theories overlap in many areas, they also diverge in their explanations of mental representation of quantity and cognitive domains responsible for numeric processing. Encoding Complex and Triple Code Theories both agree that stimulus formatting can largely influence both mental representation of quantity and subsequent numeric processing; however, Abstract Code Theory stipulates that regardless of stimulus format, mental representations are amodal *abstract codes* and subsequent numeric processing relies on these *abstract codes*. Triple Code and Abstract Code Theories both agree that numeric processing relies on cognitive domains specialized for processing quantity; however, Encoding Complex Theory stipulates that numeric processing relies on cognitive domains which are not modular and not unique to processing quantity. Clearly, encoding (forming mental representations) and cognitive dimensionality of numeric processing are major areas of departure for these theories.

In terms of specifying domains which may help to facilitate numeric processing, both Encoding Complex and Triple Code Theories suggest that the *language* domain (retrieving verbal information about number facts) may contribute to numeric processing. Encoding Complex Theory is perhaps the most prescriptive in specifying additional domain general contributions to numeric processing. Encoding Complex Theory suggests that working memory, domain general reasoning, and attention/inhibition are all important for successful numeric processing. Pieces of these domain general capacities are reflected in other theories of cognitive arithmetic (e.g., Abstract Code Theory mentions that working memory is of interest to numeric processing; Triple Code Theory mentions that *executive* domains involving coordinating attention are of interest to numeric processing). However, the centrality of all of these domain general capacities is made clear in Encoding Complex Theory, as well as the stipulation that they work in concert to perform a variety of problem-solving activities (i.e., that

arithmetic cognition is simply one form of problem-solving that happens to involve operating on quantities).

From a larger cognitive theoretical position, working memory, domain general reasoning, and attention/inhibition are three separate but related constructs that form the basis for *executive attention*, the ability to form and maintain mental representations of problems and problem-solving goals robust to distractions during problem-solving activities (Engle & Kane, 2003; Engle, Kane, & Tuholski, 1999; Engle & Oransky, 1999; Kane & Engle, 2002). *Executive attention* is distinct from *general intelligence*, though *executive attention* is related to the larger idea of *general intelligence* via the importance that the construct of *fluid intelligence* serves for each. *Executive attention* is thought to be carried out by distinct neural substrates in the prefrontal cortex (particularly the dorsolateral PFC), and behaviorally, is typically measured by fluid intelligence, working memory capacity, and attention/inhibition (Engle & Kane, 2003; Kane & Engle, 2002). Though the theory of *executive attention* allows for these three capacities to be distinct (i.e., to maintain distinct variance), their overlapping contributions to complex problem-solving tasks that demand sustained attention and goal maintenance in the face of distraction (i.e., their covariance) is thought to reflect the larger *executive attention* construct (Kane & Engle, 2002). Notably, for the purposes of the current study, Encoding Complex Theory views arithmetic cognition as one example of a complex problem-solving task. Thus, Encoding Complex Theory's hypothesized, joint contributions of working memory, domain general reasoning, and attention/inhibition may be best represented by the larger cognitive construct of *executive attention*.

Measurement Hypotheses for the Current Study

Given these varying theoretical accounts of arithmetic cognition, the purpose of the current study was to examine arithmetic cognition on symbolically formatted measurement instruments, with attention to potential formatting effects and possible contributions from cognitive abilities other than a *quantitative* domain that is specialized for numeric processing. Each leading theory of arithmetic cognition was used to formulate a series of measurement hypotheses, and a multitrait, multimethod methodology was used in conjunction with confirmatory factor analysis to examine each set of hypotheses. The architecture of an *arithmetic* domain(s), implications of that architecture for measuring various problem formats, and contributions of *language* and *executive attention* domains were simultaneously specified and estimated in the larger measurement models for each theory under investigation.

Method

Participants

Participants were drawn from public schools in a metropolitan school district in the Southeastern United States. During the fall of each third-grade school year, students who assented to participate and whose parents consented to participate in the study were included in assessment (and instructional intervention for the purposes of a parent study focused on testing the effects of an experimental instructional program for mathematics problem solving and its cognitive correlates; see e.g., Fuchs et al., 2008). An initial 2,023 students across 120 classrooms had consent to participate in the parent study. A subset of N = 1,320 children were randomly selected for full participation in the parent study and received the full testing battery (including screening measures, the full mathematics battery, cognitive measures, and demographic reports from teachers). A final sample of 1,314 children was selected for the current study.

The final sample had a mean age of 103.24 months (SD = 5.41, range = 89–142), was approximately 50% female (n = 661 females, n = 652 males), and was ethnically and racially diverse (43% African American, 40% White, 10% Hispanic, 1% Kurdish, 4% other not specified, and 1% missing). Approximately 56% of the children in the sample qualified for free or reduced lunch. Teachers reported that approximately 5% of the children in the parent study sample were receiving special education services. Of those 67 children whose teachers reported receiving special education services, most were receiving services for learning disabilities (N = 22), speech/hearing/language (N = 21), attention-deficit-hyperactivity disorder (ADHD; N = 7), or giftedness (N = 4).

Procedures

The parent study was designed to sample four, consecutive cohorts of third grade students, following each cohort for three academic years spanning from the fall of third grade until the spring of fifth grade. The current analysis, however, relies only on baseline testing for each of these four cohorts of students (i.e., all measures in the current study were administered before the sample cohorts received any intervention in the parent study). Table 1 displays cohort sampling information.

During September and October of each year of the study, (a) a demographic questionnaire was completed by teachers, (b) students' mathematical skills were assessed in three sessions lasting 30–60 min each, and (c) students' cognitive abilities were assessed in two sessions lasting 45 min each. Total testing span from first assessment to last was approximately 1 month.

The cognitive battery (described below) was administered individually by trained assessment professionals in quiet testing locations within schools. Standardized mathematics assessments were administered using recommended test developer procedures, and nonstandardized mathematics assessments were administered to students using a whole classroom assessment methodology. Students received individual stimulus papers and pencils. Trained assessment professionals read questions aloud while students followed along on their own paper copies. Students were given time to respond to each question, and the next question was not administered until all students or all but two students had put their pencils down. Students were not permitted to communicate answers or disrupt the testing of the whole class. Table 2 presents descriptive statistics for mathematics, language, and executive attention measures.

Measures

For each measure, correct items were scored "1," and incorrect items as "0" unless otherwise noted. Total raw score was the number of correct items (or partially correct items in noted in-

Table 1
Cohort Measurement Information

Measures	Cohort 1 received	Cohort 2 received	Cohort 3 received	Cohort 4 received
Mathematics measures				
WJ-III applied problems	X	X	X	X
Single digit story problems	X	X	X	X
Vanderbilt complex story problems				X
Basic facts addition	X	X	X	X
Basic facts subtraction	X	X	X	X
WRAT written arithmetic		X	X	X
Test of computational fluency	X	X	X	X
Double digit addition	X			
Double digit subtraction	X			
Double digit addition estimation	X			
Double digit subtraction estimation	X			
Language measures				
WASI vocabulary	X	X	X	X
WDRB listening comprehension	X	X	X	X
TOLD grammatic closure	X	X	X	X
Executive attention measures				
SWAN Teacher Survey	X	X	X	X
WMTB listening recall	X	X	X	X
WJ-III numbers reversed	X	X	X	X
WASI matrix reasoning	X	X	X	X
WJ-III concept formation	X	X	X	X
Cohort sampling information	N = 491 students N = 30 classes N = 7 schools	N = 485 students N = 30 classes N = 8 schools	N = 452 students N = 29 classes N = 8 schools	N = 531 students N = 31 classes N = 9 schools
Total sample for the current study	N = 1,959 students N = 120 classrooms (classrooms do not overlap) N = 16 schools (schools do overlap across cohorts)			

Note. WASI = Wechsler Abbreviated Scale of Intelligence; WDRB = Woodcock Diagnostic Reading Battery; TOLD = Test of Language Development; WMTB = Working Memory Test Battery.

stances), and this score was used in analyses. We report model-based reliability, in the form of R^2.

Mathematics achievement measures with language formatting.

WJ III Applied Problems. This measure consists of 60 orally presented word problems designed to represent every day, practical math problems (McGrew & Woodcock, 2001). Items require examinees to count, perform simple arithmetic operations, tell time, tell temperature, or problem-solve by eliminating extraneous information (McGrew & Woodcock, 2001). It is important to note that some of the WJ III Applied Problems items do not represent the traditional word problems that students typically encounter in school curricula. These items represent a mixture of traditional word problems and applied problems.

Single digit story problems. This measure consists of 14 word problems (adapted from Jordan & Hanich, 2000), read aloud while students follow along on their own written copies. Each item could be solved in one step with sums or minuends of 9 or less.

Complex story problems. This measure consists of 18 word problems, read aloud while students follow along on their own written copies (Fuchs, Hamlett, & Powell, 2003). Each item involves one to four steps for solution. Nine items are more complex and require students to eliminate extraneous information from the problem, solve problems involving novel contexts using real-world information and their own problem-solving experiences, and apply information and solutions generated in previous segments of the complex problem. Students could earn a total of 2 points per item, 1 point for correctly calculating intermediate steps in the problem, and 1 point for correctly labeling the final answer.

Mathematics achievement measures with Arabic numeral formatting.

Basic facts addition. This measure consists of 25 addition fact items (Fuchs et al., 2003). Each item involves addends of 9 or less and sums of 12 or less. Students are provided with the stimulus paper and a pencil and permitted one minute to complete as many items as possible.

Basic facts subtraction. This measure consists of 25 subtraction fact items (Fuchs et al., 2003). Each item involves minuends of 18 or less and answers of 12 or less. Students are provided with the stimulus paper and a pencil and were permitted one minute to complete as many items as possible.

WRAT Written Arithmetic. The WRAT-3 Written Arithmetic subtest (Blue form; Wilkinson, 1993) consists of 40 computation problems. Students are provided a pencil and asked to produced written responses to as many items as possible within 15 min. Items contain a variety of arithmetic content including basic facts, arithmetic involving multiple operands, arithmetic operations with proportions, and reducing and evaluating algebraic expressions (Wilkinson, 1993).

Second grade computational fluency. This measure consists of 25 items and is designed for second grade addition, subtraction, number combinations, and procedural computation problems (Fuchs, Hamlett, & Fuchs, 1990). Examinees are given 3 min to complete as many problems as possible.

Table 2
Descriptive Statistics for All Measures

Measure	*N*	Mean (*SD*)	Range: Min.–Max.
Mathematics measures			
WJ-III applied problems	1,302	29.15 (4.32)	2–48
Single digit story problems	1,307	9.96 (3.46)	0–14
Vanderbilt complex story problems	324	8.31 (6.11)	0–34
Basic facts addition	1,309	11.90 (4.99)	0–25
Basic facts subtraction	1,310	6.97 (5.03)	0–25
WRAT written arithmetic	957	23.73 (2.51)	15–34
Test of computational fluency	1,312	12.07 (6.06)	0–25
Double digit addition	339	17.12 (4.24)	0–20
Double digit subtraction	339	11.51 (5.82)	0–20
Double digit addition estimation	340	8.65 (7.11)	0–20
Double digit subtraction estimation	339	6.53 (5.90)	0–20
Language measures			
WASI vocabulary	1,314	27.35 (6.45)	5–51
WDRB listening comprehension	1,302	21.12(4.29)	0–33
TOLD grammatic closure	1,303	18.78 (6.60)	0–30
Executive attention measures			
SWAN Teacher Survey	1,258	75.71 (23.47)	18–126
WMTB listening recall	1,302	9.97 (3.58)	0–63
WJ-III numbers reversed	1,302	9.37 (2.85)	1–26
WASI matrix reasoning	1,314	15.51 (6.45)	0–30
WJ-III concept formation	1,302	15.64 (7.07)	1–39

Note. WASI = Wechsler Abbreviated Scale of Intelligence; WDRB = Woodcock Diagnostic Reading Battery; TOLD = Test of Language Development; WMTB = Working Memory Test Battery.

Double digit addition. This measure consists of 20 2-digit × 2-digit addition items with and without regrouping (Fuchs et al., 2003). Students are provided a written protocol, pencil, and 5 min to complete as many problems as possible.

Double digit subtraction. This measure consists of 20 2-digit × 2-digit subtraction items with and without regrouping (Fuchs et al., 2003). Students are provided a written protocol, pencil, and 5 min to complete as many problems as possible.

Mathematics achievement measures involving estimation or analog magnitude.

Double digit estimation addition. This measure consists of 20 symbolic 2-digit × 2-digit addition items in which students are instructed to estimate answers to the nearest 10 (Fuchs et al., 2003). Examiners complete a sample problem to demonstrate estimation and to remind students that they are not computing exact answers to problems. Students are provided with a written protocol and pencil, and given 5 min to complete as many problems as possible. Exact calculated answers were scored as incorrect.

Double digit estimation subtraction. This measure consists of 20 symbolic 2-digit × 2-digit subtraction items in which students are instructed to estimate answers to the nearest 10 (Fuchs et al., 2003). Examiners complete a sample problem to demonstrate estimation and to remind students that they are not computing exact answers to problems. Students are provided with a written protocol and pencil, and given 5 min to complete as many problems as possible. Exact calculated answers were scored as incorrect.

Language measures. Three measures of language were used. Language is commonly defined an integration of form, use, and content, a combination of skills in the areas of phonology, syntax, morphology, lexical knowledge, semantics, pragmatics, and prosody (Bloom & Lahey, 1978). Among these possible indicators of language ability, it appears that capturing listening comprehension, vocabulary knowledge, and grammatical comprehension may be essential for accurately measuring language ability (Carroll, 1993), and thus, for the purpose of this analysis, these key components of language ability were the focus of measurement.

WASI Vocabulary. The Vocabulary subtest of the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) consists of 42 items, assessing expressive vocabulary. The initial four items require students to view a picture display and provide a verbal label for the object in each picture. Remaining items require students to provide definitions for vocabulary prompts given by examiners. Responses to all items were scored "0" if incorrect, "1" if partially correct, or "2" if the targeted response was present (Wechsler, 1999).

WDRB Listening Comprehension. The Listening Comprehension subtest of the Woodcock Diagnostic Reading Battery (WDRB; Woodcock, 1997) consists of 38 sentences or passages, read aloud to examinees who are then prompted to supply the missing word at the end of each prompt. Initial items require students to complete simple verbal analogies and word associations, and as the test continues, items become more complex and require students to discern implications of the passages they have just heard (Woodcock, 1997).

TOLD Grammatic Closure. The Grammatic Closure subtest of the Test of Language Development (TOLD-Revised edition; Newcomer & Hammill, 1988) consists of 30 sentences, assessing ability to recognize, understand, and express English morphology. Students are prompted with a sentence that is missing a word and respond verbally to supply the missing word and complete the sentence (Newcomer & Hammill, 1988).

Executive attention measures. Five measures of executive attention were used. These measures emphasize the three, domain general abilities whose coordination is theorized to allow maintenance of mental representations of problems, attention, and problem-solving goals, in the face of distraction during problem-solving activities: working memory, attention/inhibition, and fluid intelligence or inductive reasoning.

SWAN. The SWAN (Swanson et al., 2012) is a teacher survey with 18 items measuring attention, inhibition, and self-regulation. This instrument is used to measure the inattentive behavior, distractibility, impulsivity, and hyperactivity characteristic of ADHD while also capturing the normal distribution of nonclinical behavior. On the first nine items, teachers rate students for various types of inattentive behavior and distractibility; on the next nine items, teachers rate students for various types of impulsive and hyperactive behaviors. Teachers respond on a 7-point Likert-type scale (7 = *far above average*, 6 = *above average*, 5 = *slightly above average*, 4 = *average*, 3 = *slightly below average*, 2 = *below average*, 1 = *far below average*).

WMTB Listening Recall. The Listening Recall subtest of the Working Memory Test Battery for Children (WMTB-C; Pickering & Gathercole, 2001) consists of sequences of sentences, assessing verbal working memory. Examiners read aloud a series of short sentences to students. After listening to each sentence, the student evaluates the sentence as true or false. Finally, after evaluating all of the sentences in a trial, the student is asked to recall, in order,

the last word of each sentence in the trial (Pickering & Gathercole, 2001).

WJ III Numbers Reversed. The Numbers Reversed subtest of the WJ-III (Test of Cognitive Abilities; Woodcock, McGrew, & Mather, 2001) consists of 30 items, assessing working memory. On each item, students listen to orally presented, random spans of digits, and upon completion of the span, students are prompted to orally list the digits they have just heard in reversed order. As students progress through the test, digit spans increase, ranging from two to eight digits (McGrew & Woodcock, 2001).

WASI Matrix Reasoning. The Matrix Reasoning subtest of the WASI is designed to measure nonverbal problem-solving or induction (Wechsler, 1999). This assessment requires examinees to view visual displays of matrices from which a section is missing and to use pattern completion, classification, analogy, and serial reasoning to induct the rule in the matrix and predict the next item in the sequence. Examinees complete the matrix using one of five possible response choices from a multiple choice array beneath the matrix prompt. Responses are identified verbally or with pointing (Wechsler, 1999).

WJ III Concept Formation. The Concept Formation subtest of the WJ-III (Test of Cognitive Abilities; Woodcock et al., 2001) consists of 40 items, assessing fluid intelligence and induction. On each item, students are shown illustrations that demonstrate instances and noninstances of a concept and are asked to identify the rules for concepts by inducting or inferring the rules (McGrew & Woodcock, 2001).

Design

The full mathematics assessment battery involved 11 measures total, and therefore, the mathematics assessments also were delivered using a planned missing design such that not all measures were administered to the random subset of children selected to receive the full battery every year of the study (for more information on planned missing designs, see e.g., Graham, Hofer, & MacKinnon, 1996). Because of the planned missingness inherent in this design, cohorts which have unavailable data on certain measures are assumed to have data that are missing completely at random, or MCAR.

Results

Planned analyses were executed in two phases of model testing. Phase one began by examining measurement models for arithmetic measures using confirmatory factor analysis with maximum likelihood estimation in MPlus 7 (Muthén & Muthén, 2012). Next measurement models for *language* and *executive attention* were examined using confirmatory factor analysis with maximum likelihood estimation in MPlus 7 (Muthén & Muthén, 2012). Phase two examined full measurement models, incorporating all constructs of interest (*mathematics*, *language*, and *executive attention* as outlined in the introduction). Missing data were estimated using full information maximum likelihood estimation (see e.g., Enders & Bandalos, 2001) in MPlus 7 (Muthén & Muthén, 2012). Note that because hypothesized model testing was extensive and included examination of 11 models, full model results are presented only for a select few of the tested models. The model results presented in text are highlighted because of their relevance to the current study's overall conclusions about the structure of arithmetic cognition, including possible formatting effects and domain specificity. However, full model testing results, including standardized and unstandardized factor loadings and intercepts as well as indicator residuals and corresponding commonalities, are available in the supplementary materials for this article.

Phase 1: Measurement Models for Arithmetic, Language, and Executive Attention

This phase of model testing began with an examination of the arithmetic portions of measurement for each of the four theories considered in this study. Figure 1 displays diagrams for each model tested, global fit statistics (exact and approximate), and completely standardized indicator factor loadings. The *abstract semantic representations* measurement model tested the extent to which the 11 mathematics indicators measure a unitary, underlying, common form of mental representation upon which all factors of numeric processing operate, in predicting mathematics outcomes. The *seemingly modular encoding complex* model tested the extent to which 11 arithmetic indicators measure a unitary, underlying, encoding complex factor, which appears to be modular with practice. It should be noted that this factor is being called "*seemingly-modular encoding complex*" here, but in actuality is the same measurement model as the *abstract semantic representations* measurement model because from a measurement standpoint, the same factor structure can be used to represent these hypotheses (though the implications and interpretations of that factor structure would be conceptually distinct across the two theories). This limitation of the factor analytic framework is considered in more detail in the Discussion section.

The Triple Code Theory model of arithmetic tested the extent to which arithmetic behavioral outcomes could be explained by three, latent factors with format and problem demand specific responsibilities in numeric processing, a *visual Arabic* factor (indicated by 6 mathematics measures), an *auditory verbal* factor (indicated by 3 mathematics measures), and an *analog magnitude* factor (indicated by 2 mathematics measures). The Exact versus Approximate Calculations Theory tested the extent to which arithmetic behavioral outcomes could be explained by two, latent factors, an *analog magnitude* factor (indicated by 2 mathematics measures), and an *auditory verbal* factor (indicated by 9 mathematics measures). Of these models, the Triple Code Theory model of arithmetic was an approximate good fit for the data, while the other three models of arithmetic measurement were not. These results support Triple Code Theory's specification that three, separate but mutually informed, format-specific factors predict arithmetic cognition outcomes.

Measurement models for *language* and *executive attention* also were examined during this phase of model testing. The *language* measurement model tested the extent to which three indicators (vocabulary, listening comprehension, and grammatical closures) measure a unitary, latent *language* ability. With three observed indicators, this latent *language ability* factor model is just-identified (i.e., has zero degrees of freedom), meaning that tests of global fit such as the χ^2 test of model fit, the root mean squared error of approximation (RMSEA), or the comparative fit index (CFI) are trivial (Brown, 2006). Though global fit could not be examined for this model, factor loadings indicated that these three

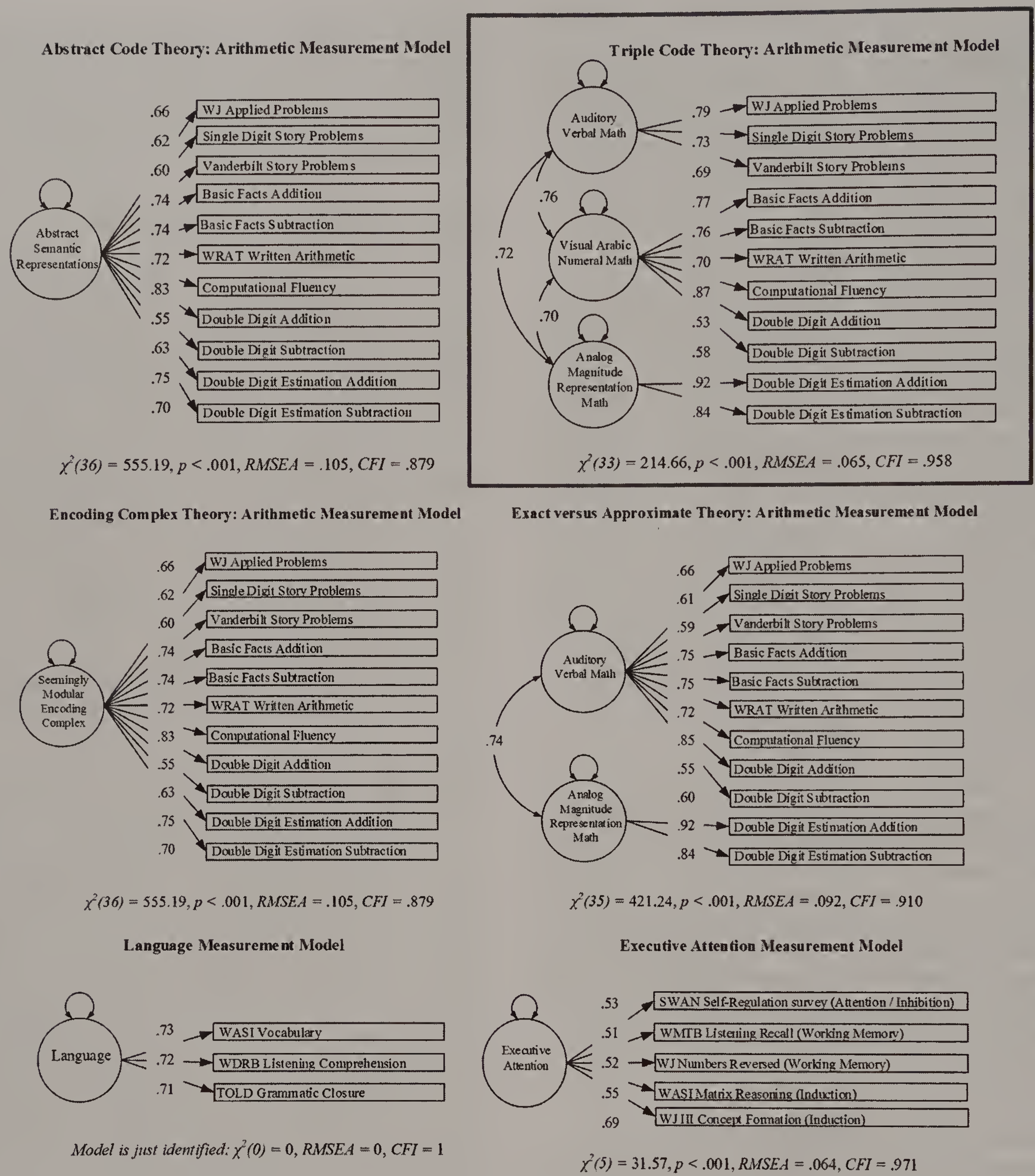


Figure 1. Summary of Phase 1 testing. The highlighted Triple Code Theory model represented the best fitting model of arithmetic measurement for this phase of testing. See the online article for the color version of this figure.

indicators were reasonable measures of the same underlying dimension.

The *executive attention* measurement model tested the extent to which five indicators measure a unitary, underlying, *executive attention* ability which was hypothesized to be indicated by a measure of attention and inhibition (the SWAN teacher survey), two measures of verbal working memory (the WJ-III Numbers Reversed and the WMTB-C Listening Recall subtests), and two measures of fluid intelligence and inductive reasoning (the WASI Matrix Reasoning and the WJ-III Concept Formation subtests). Though this model was an approximate good fit for the data, all indicators in this model demonstrated relatively high residuals. The *executive attention* model, though adequate for the purposes of the current study, evidenced issues of fit that could be interpreted to mean that important complexity in this construct was not being modeled with a unitary conceptualization.

These results were not surprising given that theoretically, *executive attention* is a construct that represents hierarchical overlap between the three separable abilities of *working memory*, *fluid intelligence*, and *attention/inhibition*. Their covariance represented

coordination and joint contributions to sustained attention and goal maintenance during problem-solving. As the current study was focused on their overlap in predicting arithmetic performances across various formats, and because the model was an approximate good fit for the data, this model of *executive attention* was ultimately retained for further testing.

Phase 2: Full Measurement Models for Each Theory

The next phase of model testing examined each of the four theories of arithmetic cognition with the inclusion of *language* and *executive attention* abilities in full measurement models. Figure 2 displays diagrams for each model tested, global fit statistics (exact and approximate), and completely standardized indicator factor loadings. Each model is briefly presented in the sections that follow.

Abstract Code model. The full measurement model of Abstract Code Theory was represented with a one factor model of *abstract semantic representation*, which at a minimum, was allowed to correlate with other cognitive domains (e.g., *language*, *executive attention*). Global fit statistics indicated that this factor

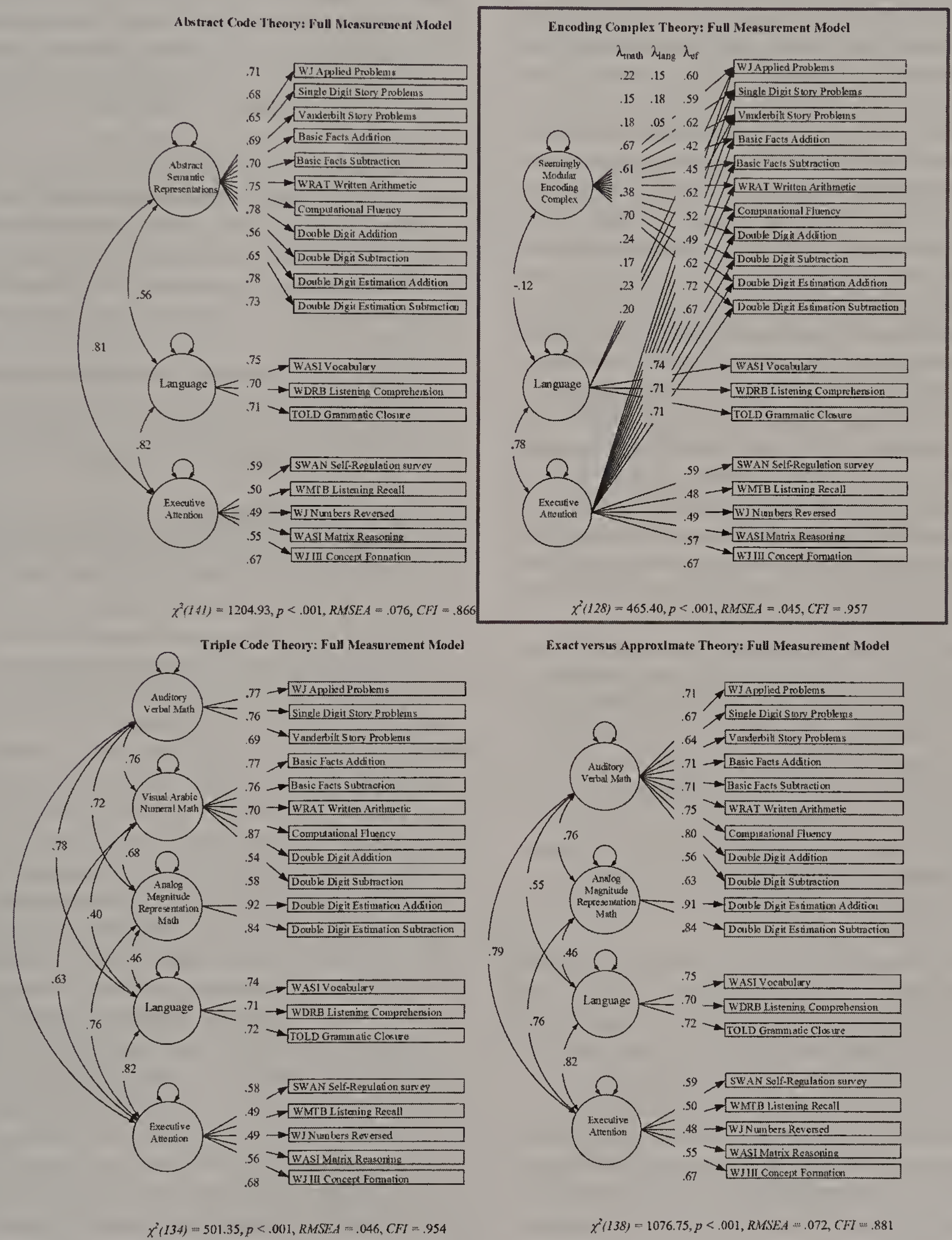


Figure 2. Summary of Phase 2 testing. The highlighted Encoding Complex Theory model represented the best fitting model of arithmetic measurement for this phase of testing. See the online article for the color version of this figure.

model was not an approximate good fit for the data, ($\chi^2(141) = 1204.93$, $p < .001$, RMSEA = .08, CFI = .87). Completely standardized factor loadings ranged from .49 to .78; indicator residual variances ranged from .39 to .77; and Model R^2 ranged from .24 to .61. Although the correlation between *language* and *abstract semantic representations* was moderate, $r = .56$, the correlations between *executive attention* and *abstract semantic representations* and *executive attention* and *language* were quite high ($r = .81$ and $r = .82$, respectively). Although both the *abstract semantic representations* and *executive attention* measurement model results suggest that both of these factors are contributing to the model misfit for the Abstract Code Theory full measurement model, the patterns of factor correlation would suggest that the relationships between *executive attention* and other constructs in the model were also important sources of model misspecification.

Encoding Complex model. The full measurement model of Encoding Complex Theory allowed *language* and *executive attention* to directly predict arithmetic outcomes along with the *seemingly modular encoding complex for arithmetic*. *Executive attention* was allowed to predict arithmetic behavioral outcomes across various formats and problem demands; however, *language* was allowed to predict arithmetic behavioral outcomes for language-formatted problems. Global fit statistics indicated that this factor model was an approximate good fit for the data, ($\chi^2(128) = 465.40$, $p < .001$, RMSEA = .05, CFI = .96). Completely standardized factor loadings ranged from .05 (nonsignificant) to .74; indicator residual variances ranged from .24 to .77; and Model R^2 ranged from .23 to .76. As mentioned in the *executive attention* measurement model results, the residuals for this factor were high. However, *executive attention* was a significant and salient predictor of all arithmetic outcomes, and *language* was a significant predictor of WJ Applied Problems and Single Digit Story Problems, though these loadings were quite low.

Allowing for direct prediction of arithmetic outcomes by *executive attention* and *language* left little unique predictive power for the *seemingly modular encoding complex*; however, each arithmetic outcome was still significantly predicted by something other than *executive attention* and *language* (represented here by the *seemingly modular encoding complex*). Three outcomes in particular (Basic Facts Addition, Basic Facts Subtraction, and Computational Fluency, all of which were formatted with Arabic numerals and involved relatively small problem sizes) had high *encoding complex* factor loadings despite the addition of *executive attention* as a predictor.

Because *executive attention* was a direct predictor of arithmetic outcomes in this model, the correlation between *executive attention* and the *seemingly modular encoding complex* was restricted to zero for the purpose of model specification. The correlation between *executive attention* and *language* was large and positive, $r = .78$; however, the correlation between *language* and the *encoding complex* was small and negative, $r = -.12$. These results indicate that although *language* is a small but significant predictor of outcomes in language-formatted problems, it is not a predictor of outcomes in Arabic numeral formatted problems or estimation problems.

Triple Code model. The full measurement model for Triple Code Theory allowed a latent *language* factor and an *executive attention* factor to correlate with the *auditory verbal*, *visual*, and *analog magnitude* factors of Triple Code Theory. Global fit statistics indicated that this factor model was an approximate good fit for the data, ($\chi^2(134) = 501.35$, $p < .001$, RMSEA = .05, CFI = .95). Completely standardized factor loadings ranged from .49 to .92; indicator residual variances ranged from .16 to .76; and Model R^2 ranged from .24 to .84. As in the Triple Code Theory arithmetic measurement model, the arithmetic portion of this full model was very strong. Completely standardized factor loadings ranged from .54 to .92, and factor correlations for this portion of the model ranged from $r = .68$ to $r = .76$, indicating that each of Triple Code Theory's arithmetic cognition factors were separable but highly related. Again the *executive attention* measurement model results demonstrated high residuals. However, *executive attention* factor loadings indicated that the selected outcomes were all significant and salient indicators of this factor. The *executive attention* factor correlated highly with all other factors in the Triple Code Theory model.

The addition of *executive attention* raised some structural questions for the arithmetic portion of the Triple Code Theory model. Specifically, the relationship between *executive attention* and the *auditory verbal factor* was nearly at singularity, $r = .94$, and the relationship between *language* and the *auditory verbal* factor was also quite high, $r = .78$. Taken together, these results indicate that (a) problem formatting should be explicitly accounted for in modeling arithmetic outcomes, (b) *executive attention* and *language* may both play important roles in facilitating arithmetic cognition across various problem formats, but (c) language-formatted items in particular may be predicted by domains other than a specialized quantitative domain.

Exact versus approximate model. The full measurement model of Exact versus Approximate Calculations Theory allowed a latent *language* factor and an *executive attention* factor to correlate with the both the *analogical magnitude representation* factor (predicting tasks requiring approximate calculations) and an *auditory verbal* factor (predicting tasks requiring exact calculations). Global fit statistics indicated that this factor model was not an approximate good fit for the data, ($\chi^2(138) = 1076.75$, $p < .001$, RMSEA = .07, CFI = .88). Completely standardized factor loadings ranged from .48 to .91; indicator residual variances ranged from .17 to .77; and Model R^2 ranged from .23 to .83. Although both the *exact versus approximate calculations* and *executive attention* measurement model results suggest that all of these factors are contributing to the model misfit for the Exact versus Approximate Calculations full measurement model, the patterns of factor correlation would suggest that the relationships between *executive attention* and other constructs in the model may also be important sources of model misspecification.

Executive attention correlated significantly and strongly with all other factors in the model, indicating that executive systems of control may indeed play a role in facilitating both exact and approximate calculations. *Language*, however, correlated only moderately with the *auditory verbal* and *analog magnitude* factors, but it correlated highly with *executive attention*. Taken together, this pattern of correlations would seem to suggest that *language* is separable from traits predicting arithmetic outcomes across exact and approximate problem demands, which are in turn both highly related and separate from each other (*auditory verbal* and *analog magnitude* factors correlated at $r = .76$).

Summary of Model Testing Results

Results from the arithmetic only measurement models indicated that the Triple Code Theory model of arithmetic was the best fitting model; however, the Triple Code Theory full measurement model displayed some structural problems, namely a correlation between *executive attention* and the *auditory verbal* factor that was near singularity and very high correlations between *executive attention* and the other factors of arithmetic in the model.

Conversely, results from the Encoding Complex full measurement model indicated that this model of arithmetic (and its relationships with other cognitive domains) was the best fitting model; however, the architecture for arithmetic in the Encoding Complex Theory model was unidimensional, and results from the arithmetic only measurement models indicated that a unidimensional arithmetic was not a good fit for the data.

Given that (a) a three-factor model of arithmetic presented by Triple Code Theory was an excellent fit for the data and (b) a direct prediction of *executive attention* and *language* on math outcomes presented by Encoding Complex Theory was an excellent fit for the data, results supported both Encoding Complex Theory and Triple Code Theory. Thus, a final, unplanned, post hoc model, incorporating key measurement hypotheses of each theory, was examined.

Post Hoc Testing: Hybrid Full Measurement Model

The post hoc model represents the three-factor arithmetic (only) portion of Triple Code Theory with Encoding Complex Theory's specification that *executive attention* could be a direct predictor of all arithmetic outcomes and that *language* could be a direct predictor of outcomes on language-formatted arithmetic problems. A *visual Arabic* factor processes digital input and output as well as multidigit operations. An *auditory verbal* factor processes simple mathematical facts, language-based input and output, and language-based memory for numbers. An *analog magnitude* factor processes semantic information for number and is responsible for performing comparison, estimation, approximate calculation, and subitizing tasks across various formats of input and output. Transcoding allows for these factors to inform one another directly during numeric processing tasks. The post hoc hybrid model represents each of these factors as a latent factor and transcoding as the correlation between these factors.

Global fit statistics indicated that this factor model was an approximate excellent fit for the data, ($\chi^2(124) = 327.82$, $p < .001$, RMSEA = .04, CFI = .97). Across outcomes, completely standardized factor loadings ranged from −.003 (nonsignificant) to .74; indicator residual variances ranged from .22 to .76; and Model R^2 ranged from .24 to .78. As mentioned in the *executive attention* measurement model results, the residuals for this factor were high and among the highest in the model. Table 3 presents completely standardized results for the Hybrid full measurement model. Figure 3 displays a model schematic with completely standardized factor loadings, indicator residuals, and latent factor correlations.

More specifically, for the arithmetic outcomes across the three Triple Code factors, completely standardized factor loadings ranged from .18 to .70. All of these loadings were significant, but only the factor loadings for the following five arithmetic outcomes were salient: Basic Facts Addition, Basic Facts Subtraction, and Computational Fluency (all Arabic numeral formatted and all involving relatively small problem sizes), as well as Double Digit Estimation Addition and Double Digit Estimation Subtraction (both involving estimation). For the *language* outcomes, completely standardized factor loadings were all significant and sa-

Table 3
Post Hoc Hybrid Full Measurement Model Completely Standardized CFA Results

Indicator	Intercepts (*SE*)	Factor loadings (*SE*) by Factor: Auditory verbal	Visual Arabic number	Analog magnitude	Language	Executive attention	Residual variance	R^2
Arithmetic measures								
Applied problems	6.71 (.14)	.32 (.05)			.04 (.06)^{NS}	.67 (.06)	.40	.60
Single digit story problems	2.89 (.06)	.21 (.04)			.09 (.06)^{NS}	.65 (.06)	.43	.57
Complex story problems	1.33 (.07)	.28 (.08)			−.003 (.13)^{NS}	.64 (.12)	.52	.49
Basic facts addition	2.39 (.05)		.67 (.02)			.42 (.03)	.37	.63
Basic facts subtraction	1.39 (.04)		.61 (.02)			.46 (.03)	.42	.58
Written arithmetic	9.42 (.21)		.38 (.03)			.62 (.02)	.47	.53
Computational fluency	1.99 (.05)		.70 (.02)			.52 (.03)	.23	.77
Double digit addition	4.04 (.16)		.24 (.06)			.48 (.04)	.71	.29
Double digit subtraction	1.98 (.09)		.18 (.05)			.61 (.04)	.60	.40
Double digit estimation addition	1.21 (.06)			.53 (.06)		.71 (.03)	.22	.78
Double digit estimation subtraction	1.09 (.06)			.60 (.06)		.64 (.04)	.23	.77
Language measures								
Vocabulary	4.24 (.09)				.74 (.02)		.45	.55
Listening comprehension	4.88 (.10)				.71 (.02)		.50	.50
Grammatic closure	2.82 (.06)				.72 (.02)		.49	.51
Executive attention measures								
Attention	3.20 (.07)					.59 (.02)	.65	.35
Listening Recall	2.78 (.06)					.49 (.02)	.76	.24
Numbers reversed	3.28 (.07)					.49 (.02)	.76	.24
Matrix reasoning	2.41 (.05)					.57 (.02)	.68	.32
Concept formation	2.20 (.05)					.67 (.02)	.55	.45

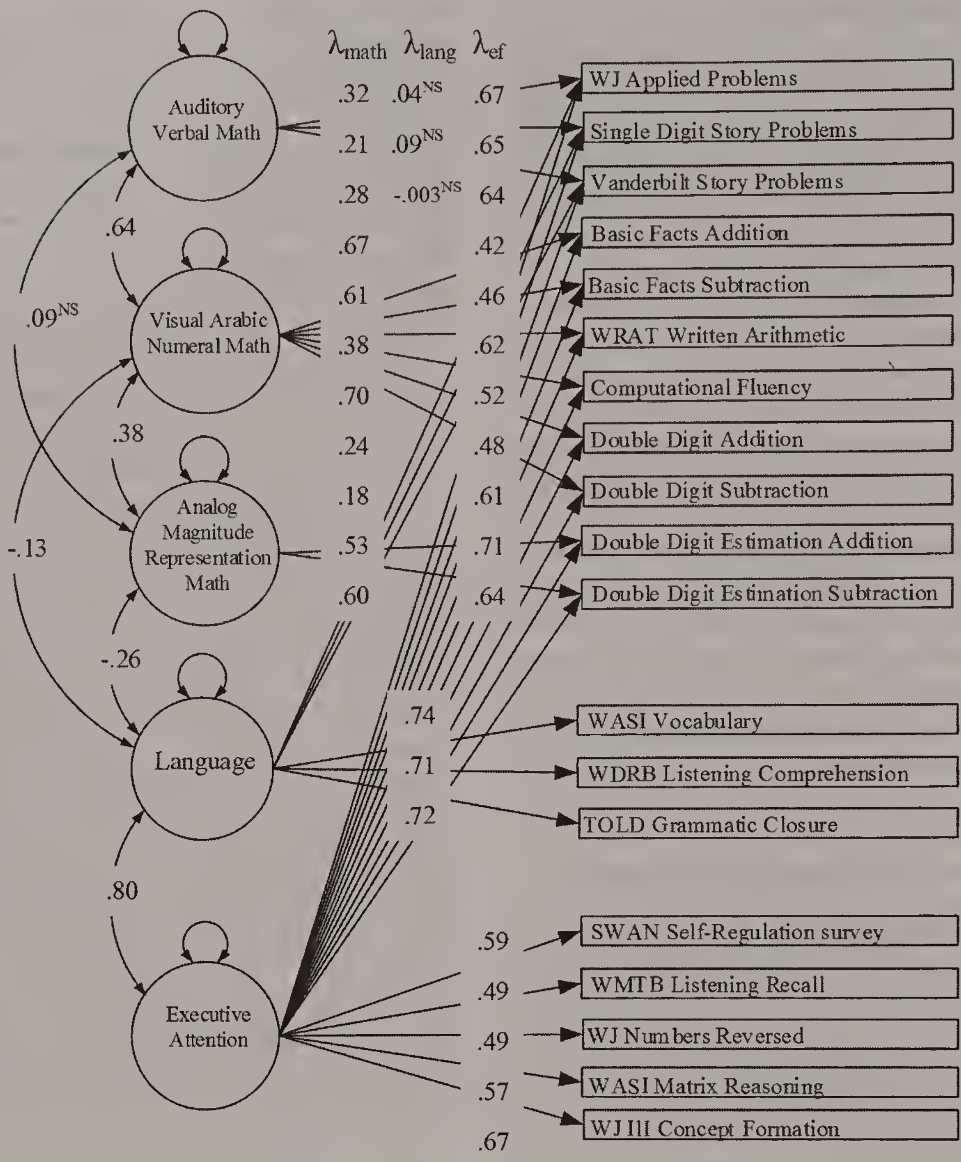


Figure 3. Hybrid final arithmetic measurement model: Triple code arithmetic structure and encoding complex measurement model.

lient, ranging from .71 to .74; however, none of the language-formatted arithmetic outcomes were significant indicators of *language*, meaning that the *auditory verbal* factor is distinct from *language*. For the *executive attention* outcomes, completely standardized factor loadings were all significant and salient, ranging from .49 to .67. The arithmetic outcomes were all significantly and saliently indicated by the *executive attention* factor. Completely standardized factor loadings ranged from .42 to .70, and they were lowest for the three aforementioned Arabic numeral formatted/small problem size outcomes.

Allowing for direct prediction of arithmetic outcomes by *executive attention* and *language* left little unique predictive power for the three Triple Code Theory factors of arithmetic; however, all of the arithmetic outcomes were still significantly predicted by its corresponding Triple Code Theory factor. This pattern of results indicates that something other than *executive attention* and *language* (represented here by the *visual Arabic number form* factor, *auditory verbal* factor, and *analog magnitude* factor) was predicting performance for each of these problem formats or analogical magnitude demands. The *auditory verbal* factor loadings were all particularly low with *executive attention* in the model, which would seem to indicate that language-formatted problems, in particular, are largely *executive attention* tasks.

Because *executive attention* was a direct predictor of arithmetic outcomes in this model, the correlations between *executive attention* and the *visual Arabic number form* factor, the *auditory verbal* factor, and the *analog magnitude* factor were restricted to zero for the purpose of model specification. Similarly, the correlation between *language* and the *auditory verbal* factor was also restricted to zero. The correlation between *executive attention* and *language* was large and positive, $r = .80$; however, the correlations between *language* and both the *visual Arabic number form* and *analog magnitude* factors were small and negative, $r = -.13$ and $r = -.26$, respectively. Among the Triple Code Theory factors, *auditory verbal* arithmetic and *visual Arabic number form* arithmetic were moderately and positively related, $r = .64$, and *analog magnitude* arithmetic and *visual Arabic number form* arithmetic were slightly and positively related, $r = .38$. However, the *audi-*

tory verbal and *analog magnitude* factors were not significantly related.

This model was an approximate good fit for the data, and χ^2 difference testing indicated that this model significantly improved fit as compared with all other full measurement models tested (see Table 4). This model represented a synthesis of hypotheses from two theories of arithmetic cognition that were supported by patterns of results from all model testing, and as such, this model was ultimately retained as the most parsimonious presentation of results.

Discussion

The purpose of this study was to evaluate the effects of item formatting and to explore the possibility that *language* and *executive systems of control* contribute to solving various formats of arithmetic problems. This research was approached using multitrait, multimethod data, and confirmatory factor analysis. Four leading theories of arithmetic cognition were used to guide measurement hypotheses about the (a) structure of mathematics abilities involved in arithmetic cognition, (b) roles of symbolic problem formatting (language vs. Arabic numeral formats) and calculation demands (exact vs. approximate) in predicting arithmetic outcomes, and (c) possible contributions of *language* and *executive attention* in predicting arithmetic outcomes.

Summary of Major Findings

As predicted by Triple Code Theory, the structure of arithmetic cognition was best supported by several latent factors of *quantitative* ability with specialization for particular formats and problem demands. Put in terms of psychometric theory, similarly formatted problems displayed common method variance that was explained by unique factors of arithmetic ability. An *auditory verbal* factor was largely responsible for problems that were language-formatted. A *visual Arabic number form* factor was largely responsible for problems that were formatted with Arabic numerals. An *analog magnitude* factor was largely responsible for problems that involved estimation across formats. This three-factor architecture of arithmetic cognition was valuable for explaining arithmetic outcomes across the models tested in the current study.

Abstract Code Theory's stipulation that *abstract semantic codes* predict arithmetic outcomes across various formats of problem was not supported, nor was a specification of Encoding Complex Theory in which a unitary, *seemingly modular encoding complex* predicts arithmetic outcomes across formats. Exact versus Approximate Calculations Theory's specification that exact and approximate problem demands would be predicted by separable cognitive architectures was somewhat supported. Among calculation demands, exact and approximate calculations were distinct but related; however, within exact problems, those problems with language formatting were separable from problems with Arabic numeral formatting.

As predicted by Encoding Complex Theory, *executive attention* was a major predictor of all arithmetic outcomes. The inclusion of *executive attention* as a direct predictor of arithmetic outcomes overwhelmed the arithmetic-only models of cognition. Little variance remained for factors of arithmetic cognition to explain; however, each retained some unique predictive value.

An interesting find was that *executive attention* left no predictive value for *language* on language-formatted problems. Language-formatted problems were explained mostly by *executive attention* and somewhat by the *auditory verbal* factor of arithmetic, and *language* evidenced a negative relationship with Arabic numeral formatted problems and estimation problems. This outcome suggested that *language* was not directly contributing to arithmetic cognition. However, the lingering, large correlation between *language* and *executive attention* suggested that *language* had some role to play in arithmetic cognition. Taken together, these findings raise questions about the possibility that *language* may play a facilitative role in reasoning, particularly for language-formatted problems.

Explaining the relationship between *language* ability and *executive attention* in a theoretical model of arithmetic cognition will be a challenge for future research. Because *language* was not positively associated with factors of arithmetic, because *language* was not a direct predictor of language-formatted arithmetic, and because *executive attention* was a direct predictor of arithmetic outcomes across factors of cognition, this study suggests that *language* may play an indirect role in helping *executive systems of control* to predict arithmetic outcomes.

Table 4
Summary of Model Testing Results

Models tested	χ^2	*df*	*p*	CFI	RMSEA	Note
Initial measurement models						
Abstract Code Arithmetic	555.19	36	<.001	.88	.11	
Encoding Complex Arithmetic	555.19	36	<.001	.88	.11	Structurally Identical to Abstract Code Arithmetic
Triple Code Arithmetic	214.66	33	<.001	.96	.07	
Exact vs. Approximate Arithmetic	421.24	35	<.001	.91	.09	
Language	.00	0	N/A	1.00	.00	Model is just-identified
Executive Attention	31.57	5	<.001	.97	.06	
Full measurement models						
Abstract Code Theory	1204.93	141	<.001	.87	.08	$\Delta\chi^2(17) = 877.11$, $p < .001$
Encoding Complex Theory	465.40	128	<.001	.96	.05	$\Delta\chi^2(4) = 137.58$, $p < .001$
Triple Code Theory	501.35	134	<.001	.95	.05	$\Delta\chi^2(10) = 173.53$, $p < .001$
Exact vs. Approximate Theory	1076.75	138	<.001	.88	.08	$\Delta\chi^2(14) = 748.93$, $p < .001$
Post hoc hybrid	327.82	124	<.001	.97	.04	Baseline model for χ^2 difference testing

Note. CFI = comparative fit index; RMSEA = root mean square error of approximation.

Several theories have implicated *language* as a facilitator of systems of executive control. Most often, this relationship has been conceptualized in terms of the construct of internal speech, also called self-directed speech or private speech. Internal speech is not directed socially toward communication partners other than the self, for the purpose of facilitating cognition and behavioral control (see e.g., Berk, 1999). In Baddeley's (see e.g., Baddeley, 1992, 2000; Baddeley & Logie, 1999) model of working memory, internal speech may play a critical role in helping to maintain mental representations of stimuli via an articulatory rehearsal system. In Barkley's (1997) model of self-regulation, internal speech helps to regulate inhibitory control by guiding rule-governed behaviors and self-evaluation during problem-solving. Similarly, in Zelazo's (see e.g., Zelazo & Frye, 1998) model of problem-solving, self-directed, internal speech plays a crucial role in problem-solving, particularly during planning, inhibition, and evaluation.

Measuring internal speech may require methods that use careful behavioral observation and self-reporting during and after the performance of problem-solving tasks (Berk, 1999). Though this was beyond the scope of the current study, future research should investigate the construct of internal speech as an indirect predictor of arithmetic problem-solving.

The addition of *executive attention* as a direct predictor of arithmetic outcomes not only impacted the relations between factors of arithmetic and *language*, but also *executive attention* impacted the relationships between the three factors of arithmetic cognition. Although the three factors of Triple Code Theory evidenced a pattern of strong, positive relationships when modeled in isolation, this was no longer true when *executive attention* was explicitly modeled. Problems involving exact calculations remained highly related across language formats (on the *auditory verbal* factor) and Arabic numeral formats (on the *visual Arabic number form* factor); however, the relationships of these factors with the *analog magnitude* factor changed when *executive attention* was included. With explicit modeling of *executive attention* in arithmetic outcomes, the *visual Arabic number form* factor was only slightly related to the *analog magnitude* factor, and the *auditory verbal* factor was no longer related to the *analog magnitude* factor.

These correlations represented Triple Code Theory's specification of transcoding, or direct communication between factors of arithmetic cognition during numeric processing, and it is this notion of transcoding that allows Triple Code Theory's arithmetic factors to avoid necessarily communicating via *abstract semantic codes*. Though direct communication between Triple Code Theory's factors is assumed during numeric processing, only the *analog magnitude* factor is hypothesized to contain semantic information about number. These findings suggest that when the role of *executive attention* in arithmetic cognition is directly modeled, transcoding with the *analog magnitude* factor may be minimal or nonexistent. Perhaps numeric processing for problems involving language-formats, Arabic numeral formats, multidigit operations, and language-based memory for numbers relies more heavily on temporarily maintained mental representations of problems in a coordinated system of *executive attention* than it does on semantic information about number.

Implications for Measuring Arithmetic

The findings from the current study raise important questions about the inferences that can confidently be made from various formats of arithmetic tests. The assumption that all assessments that involve arithmetic are inherently measures of *arithmetic* ability, and only *arithmetic* ability, may not be warranted. Features of problem formatting and problem demands may influence the extent to which measurement instruments capture arithmetic ability, and even when measures appear to reliably and validly capture *arithmetic* skill, they may also be measures of *executive systems of control*.

When attempting to measure *arithmetic* cognition, measurement formatting and problem demands are important, but all of the arithmetic outcomes in the current study were largely predicted by domain general capacities in *executive attention*. Despite the overwhelming effect of *executive attention*, several measures of *arithmetic* did retain unique predictive value that was salient. These measures either involved Arabic numeral formatting and small problem sizes or estimation problem demands. Such formats and problem demands may be promising methods of assessing *basic conceptual* competence because these types of problems remained strong predictors of *arithmetic* cognition despite the contributions of *executive attention*.

Conversely, language-formatted arithmetic items may yield results with dubious inferential value for assessing some "pure" construct of *arithmetic* cognition. Language-formatted items retained little unique predictive value once *executive attention* was added as a direct predictor of arithmetic outcomes, suggesting that language-formatted items may be mostly measures of *executive attention* and, by extension, the role of *language* ability in facilitating linguistic problem-solving. Thus, language-formatted "arithmetic items," may more accurately be labeled "linguistically formatted problem-solving tasks that involve some arithmetic."

Given prior research findings and the fact that word problems are intentionally designed to reflect real-world problem-solving experiences, it is not surprising that word problems exhibited these patterns of multidimensionality (indicated by both *executive attention* and *arithmetic* ability); however, if they are multidimensional measures of both *executive attention* and *arithmetic* abilities, they should be interpreted as such as opposed to being collapsed into interpretations of mathematical ability based on other problem formats and demands. In other words, arithmetic word problems do not appear to be measures of a "pure" arithmetic ability; they also largely appear to be measures of the ability to form and maintain mental representations of problems and problem-solving goals robust to distractions during problem-solving activities. For interpretations of examinee performances on word problems to be accurate and valid, the multidimensional nature of these problems should not be ignored, and the elements of sustained and coordinated attention that they require (i.e., not just conceptual knowledge, but also strategic and procedural competence) should be acknowledged.

Limitations and Future Directions

Adapting theories toward specific measurement hypotheses. The specificity required by the factor analytic framework is a limitation of the current project. Factor models represent abilities or commonalities between various measures, but they do not

represent processes unless a process is specifically being modeled (Carroll, 1993). Such a model would necessitate structural hypotheses among traits, with the specific allowance for traits to influence one another in the time-scale specified by the process (e.g., over seconds, minutes, days, and years). Arithmetic cognition, executive attention, and self-directed speech are processes. Inferences in the current study are limited to traits, but the relationships among traits at a single time point can give important clues about underlying processes, and factor analysis can help to answer important questions about the properties of measurements.

It is important to note that these theories of arithmetic cognition were not specified with factor analysis methodologies in mind, and so, translation into factor analytic frameworks becomes difficult when theories of arithmetic cognition do not provide explicit measurement parameters. For example, "contributions" could be conceptualized as direct predictions of latent factors, correlations between latent factors, or perhaps residual error terms. Some specific aspects of each theory lend themselves to formulations with factor models, while other aspects were not necessarily testable with this method. For example, modeling Abstract Code Theory's highly complex mechanism of numeric processing was beyond the scope of the current study. Measurement hypotheses in the current study were carefully constructed with the aim of striking a balance between faithfully representing theoretical postulates and holding the research to the methodological rigor demanded by factor analysis. Still, the measurement hypotheses for theories of arithmetic cognition are open to other interpretations. Future research should explore alternate measurement hypotheses with these theories of arithmetic cognition.

Adapting theories toward developmental hypotheses. The second limitation of the current project lies in the generalization of theory to a population at an earlier developmental stage. This project aimed to understand the arithmetic cognition of school-age children and the facets of numerical cognition that may predict their development into skilled adults. Although some theories of arithmetic cognition make specifications about growth and the ways in which one might become a skilled adult, others do not. Invariance testing (testing the hypothesis that the same cognitive architecture specified for adults can be assumed for children) is implicit in the current project. However, extant neuroimaging research has indicated that quantitative cognition of children and adolescents may be qualitatively different from that of skilled adults (e.g., Cantlon et al., 2006). Future research should examine the development of arithmetic cognition in children, adolescents, and adults utilizing a longitudinal design and explicit testing of longitudinal measurement invariance. Indeed, the measurement findings of the current study may not generalize to adolescents, adults, or even children at earlier or later developmental stages than those included in the current study, and it would not be surprising to find age-related changes in the roles of language and executive attention in various arithmetic tasks. A line of developmental research with explicit focus on longitudinal measurement invariance should inform theoretical extensions of existing theories of arithmetic cognition, addressing hypotheses about the developmental continuum of quantitative cognition and its ideal measurement.

Generalizability of symbolic formatting. Another limitation of the current project is that it is exclusively focused on numeric processing with symbolically formatted measures of arithmetic (e.g., language or Arabic numeral formats) and does not include nonsymbolically formatted measures of arithmetic (e.g., dot arrays). Although the arithmetic that children will encounter in most formalized assessment settings is symbolically formatted, developmental research on the quantitative domain is focused largely on children's performance with nonsymbolically formatted measures (e.g., Feigenson, Carey, & Hauser, 2002; Starkey & Cooper, 1980; Wynn, 1992; Xu & Spelke, 2000). Including nonsymbolically formatted measures of arithmetic in measurement batteries will be essential for establishing common scaling and examining developmental continuity in the quantitative domain, and may very well provide a more "pure" measure of numerical cognition than symbolic formats. Future research should explore arithmetic cognition, formatting effects, and domain specificity with the inclusion of nonsymbolically formatted arithmetic items in the measurement battery.

Similarly, many other aspects of item modality (e.g., timed/untimed, problem size, and number of steps required to solve a problem) as well as item content (e.g., arithmetic, algebraic reasoning, and geometry) are often controlled or varied to approximate item difficulty across various types of mathematics tasks. The purpose of the current study was to examine symbolically formatted arithmetic items with regard to theoretical specifications of the cognitive abilities involved in solving them; however, future research should examine other aspects of item modality and their effect(s) on the measurement of cognitive abilities across a variety of tasks involving differing mathematical content.

Overlap in features of item modality. Although children were instructed to use estimation to solve the double digit estimation problems, and although these items were speeded to encourage the use of the most efficient strategy for solution, it should be noted that these problems could have been solved by using the strategy of calculating the exact answer and then rounding. In other words, depending upon the strategies used by children during numeric processing, the double digit estimation problems may have been solved using a combination exact calculations and approximation. Unfortunately, the strategy usage used by children during numeric processing was beyond the scope of the current study. It is indeed probable that certain formats may be better suited for eliciting certain problem-solving strategies (e.g., nonsymbolic formats may be better suited to eliciting approximate calculation strategies; see e.g., Siegler & Shrager, 1984).

Similarly, the WJ Applied Problems subtest items are language-formatted problems designed to measure children's knowledge of and ability to solve everyday problems (e.g., telling time). These problems served different roles in different models in the current study. They were alternately loaded onto unitary factors (*abstract semantic representations* or a *seemingly modular encoding complex*), an *exact calculations* factor, and an *auditory verbal* factor. Their treatment as exact calculation items was perhaps the most questionable. Problems on the WJ Applied Problems subtest require children to produce exact answers, but they do not necessarily require children to perform exact calculations. Of the 39 problems designed for examinees who are not above average adults or who are below college-level in education, most require knowledge of numbers and operations; however, 12 items (approximately 31%) involve the production of exact answers requiring specific, applied knowledge of telling time, recognizing American money, or reading a thermometer. Thus, unfortunately, the WJ Applied

Problems subtest represented a mixture of traditional word problems and applied problems. Though this subtest was consistently significant and salient as an indicator in the models tested for the current study, generalizing of the WJ Applied Problems subtest as a test of traditional word problems requiring exact calculations is limited by the extent to which it includes applied problems.

In both the case of the double digit estimation problems and the WJ Applied Problems, issues of item-formatting overlapped with issues of item calculation demands in ways that may have led to model misfit. This caveat is particularly relevant to the exact versus approximate calculations model. This research found some support for a central tenet of exact versus approximate calculations theory; problems requiring the production of exact solutions appeared to be separable from problems requiring the production of approximate solutions. Symbolic formatting was also an important contributor to the dimensionality of arithmetic measures. However, examination of the possibility that item features may interact to predict examinee responses was beyond the scope of the current study. Future research should examine the relationship between item modality and the measurement of arithmetic cognition with explicit control in the design of item features (e.g., formatting, calculation demands), observation of children's strategy usage during numeric problem-solving, and allowances for the possibility that features of item modality may interact to predict children's responses.

Measuring and modeling executive attention. For the purposes of the current study, *executive attention* was indicated by a combination of measures of working memory, inhibition and attention, and fluid intelligence (inductive reasoning or problem-solving). These measures were combined in an a priori specified, latent factor model with the aims of (a) synthesizing important facets of *executive attention*, while (b) explicitly accounting for measurement error. However, it should be noted that across all of the full measurement models and in the *executive attention*-only measurement model, the *executive attention* factor evidenced some problems.

Although this unitary *executive attention* factor displayed good model fit in most ways, patterns of residual variance indicated that much of the complexity of these indicators was not accounted for by a single factor. The single factor called *executive attention* likely represented a hierarchical construct, which would help to explain the variance unaccounted for in fluid intelligence, working memory, and attention/inhibition indicators. For the purposes of the current study, executive attention was interpreted as an overall relationship between these key systems of control in coordinating problem-solving activity; however, future research should investigate the extent to which fluid intelligence, working memory, and attention/inhibition may make shared and unique contributions to arithmetic (e.g., a bifactor model).

Summary and Conclusions

Because this study aimed to examine the construct of arithmetic cognition by examining the formatting and dimensionality of arithmetic measures, a factor analytic framework in conjunction with a multitrait, multimethod approach was appropriate. The factor analytic framework requires explicit statements of hypotheses about model parameters, which can reveal areas of theoretical misspecification, implications of measurement techniques for construct-level inferences, as well as areas of theoretical ambiguity. Though the specificity required by a factor analytic framework can be challenging, this approach is a promising method for evaluation of the construct of arithmetic cognition and its potential measures.

Four leading theories of arithmetic cognition were used to guide measurement hypotheses in the current study. Each of the theories was designed to explain the arithmetic cognition of skilled adults. This study sought to understand the arithmetic cognition of developing children who have some formal education and exposure to arithmetic, but are still actively engaged in mathematics education. Describing a developmental continuum that links the arithmetic cognition of developing children to the cognition of skilled adults will be a crucial next step for researchers and theoreticians.

In general, results from this study provided support for both Triple Code Theory and Encoding Complex Theory, and to some extent, Exact versus Approximate Calculations Theory was also supported. As predicted by Triple Code Theory, arithmetic outcomes with language formatting, Arabic numeral formatting, and estimation demands across formats were related but distinct from one another. This finding is also compatible with Encoding Complex Theory's stipulation that formatting effects exist for arithmetic cognition. The large and enduring relationship between problems that required exact calculations (across formats) also provides support for Exact versus Approximate Calculations Theory's stipulation that exact calculation problems may draw from the same cognitive processes.

Executive attention was a direct predictor of all arithmetic outcomes. This finding is compatible with Triple Code Theory's stipulation that other cognitive domains, in particular domains responsible for coordinating visuospatial attention, may contribute to arithmetic cognition. *Executive attention* is complex, and modeling that complexity was beyond the scope of the current study; however, the facets of working memory, inhibition and attention, and induction and reasoning ability shared a unitary predictive power in explaining arithmetic.

Given the strong and enduring relationship between *executive attention* and *language* ability, and the fact that language ability was not a direct predictor of arithmetic performances, this synthesized *executive attention* may have been facilitated by *language* ability in a collaborative relationship that was beyond the scope of the current study. Future research should investigate the extent to which internal or self-directed speech may facilitate *executive attention* and indirectly predict performance on arithmetic problem-solving tasks. This pattern of results may be particularly pertinent for language-formatted arithmetic items.

Results from the current study support the growing body of literature indicating that caution should be used in interpreting the results from language-formatted arithmetic items (e.g., Abedi & Lord, 2001; Martiniello, 2009; Rhodes, Branum-Martin, Morris, Romski, & Sevcik, 2015). These items may have little construct validity as pure measures of mathematics ability, but rather appear to be largely *executive attention* tasks which also involved some arithmetic ability. Though problems formatted with Arabic numerals or involving approximate calculations were also multidimensional measures of both *executive attention* and *arithmetic* abilities, these measures retained far more predictive power for measuring *arithmetic* abilities than language-formatted problems. When *executive attention* was allowed to directly predict arithmetic outcomes on language-formatted problems, *arithmetic* abilities had either no significant or no salient pre-

dictive power. Thus, difficulties with linguistically formatted arithmetic problems likely largely indicate problems with domain general problem solving capacities, and to a lesser extent, may also indicate some domain specific arithmetic ability. Inferences about pure mathematical ability should be made with caution when they are based on results from language-formatted testing instruments, and this caution is particularly relevant to national achievement assessments that utilize language-formatting in their assessment of mathematical competence.

The notion of pure mathematical abilities raises a fundamental, ontological question for researchers and practitioners who are designing, administering, and interpreting educational assessments of basic mathematical competence: What is meant by "pure mathematical ability," and is it possible to design a symbolically formatted, educational assessment of the most basic, arithmetic skills involved in such a construct? The current study found evidence that various types of symbolically formatted arithmetic problems (a) demonstrated unique clusters of quantitative skills depending upon their designs of problem formats and calculation demands, and (b) also measured domain general executive attention ability, particularly when problems were linguistically formatted. Taken together, these findings imply that (a) different types of symbolically formatted arithmetic problems measure different constellations of skills, and (b) symbolic formats may not be appropriate for measuring some construct that is purely mathematical.

Thus, measures of arithmetic should be designed, selected, and interpreted differently with respect to their formats and problem demands. For example, if one is interested in obtaining a strong measure of conceptual number knowledge, Arabic numeral formats and problems with approximate calculation demands may be more desirable than language-formatted problems. Students experiencing difficulty with Arabic numeral formats or approximate calculation problems may be struggling with understanding concepts like place-value (i.e., the visuospatial strings of digits represented by the *visual Arabic number form*) or numerosity (i.e., the semantic understanding of a number's cardinality and ordinality represented by the *analog magnitude form*), and to a lesser extent, may also be struggling with executive attention required during problem-solving. If one is interested in understanding the roles that strategic and procedural competence play in the realm of arithmetic problem-solving, word problems may provide a more desirable measure of arithmetic. Students experiencing difficulty with word problems are likely struggling with understanding concepts like selecting appropriate strategies for problem-solving and executing the procedural steps of the strategies they select, and to a lesser extent may also be struggling with concepts like interpreting number words (i.e., the syntactic, phonological, and/or graphemic understanding of number represented by the *auditory verbal word frame*).

Regardless of the measurement technique selected for assessing arithmetic skill, researchers and practitioners should also be aware that language may be playing a crucial, indirect, and internal role in facilitating children's mathematical problem-solving. The findings of the current study suggest that language ability is not a direct predictor of arithmetic performance for many students, but rather may help students to maintain attention and coordinate problem-solving procedures. More research is needed to determine the role that internal speech may play in arithmetic problem-solving; however, the strong and enduring relationship between language and executive attention in the current study suggests that targeting executive attention or helping children to moderate their internal speech during mathematical problem-solving may be promising avenues of intervention.

References

American Association on Intellectual and Development Disabilities. (2010). *Intellectual disability: Definition, classification, and systems of supports* (11th ed.). Washington, DC: Author.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219–234. http://dx.doi.org/10.1207/S15324818AME1403_2

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ansari, D. (2007). Does the parietal cortex distinguish between "10," "ten," and ten dots? *Neuron, 53,* 165–167. http://dx.doi.org/10.1016/j.neuron.2007.01.001

Baddeley, A. (1992). Working memory. *Science, 255,* 556–559. http://dx.doi.org/10.1126/science.1736359

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 417–423. http://dx.doi.org/10.1016/s1364-6613(00)01538-2

Baddeley, A., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press.

Baddeley, A., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 28–61). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139174909.005

Ballew, H., & Cunningham, J. W. (1982). Diagnosing strengths and weaknesses of sixth-grade students in solving word problems. *Journal for Research in Mathematics Education, 13,* 202–210. http://dx.doi.org/10.2307/748556

Baranes, R., Perry, M., & Stigler, J. W. (1989). Activation of real-world knowledge in the solution of word problems. *Cognition and Instruction, 6,* 287–318. http://dx.doi.org/10.1207/s1532690xci0604_1

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin, 121,* 65–94. http://dx.doi.org/10.1037/0033-2909.121.1.65

Berk, L. E. (1999). Children's private speech: An overview of theory and the status of research. In R. M. Diaz & L. E. Berk (Eds.), *Private speech: From social interaction to self-regulation* (pp. 17–53). New York, NY: Psychology Press.

Bloom, L., & Lahey, M. (1978). *Language development and language disorders*. New York, NY: Wiley.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement, 26,* 433–450. http://dx.doi.org/10.1177/014662102237798

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology, 33,* 205–228. http://dx.doi.org/10.1080/87565640801982312

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology, 19,* 273–293. http://dx.doi.org/10.1207/S15326942DN1903_3

Campbell, J. I. (1994). Architectures for numerical cognition. *Cognition, 53,* 1–44. http://dx.doi.org/10.1016/0010-0277(94)90075-2

Campbell, J. I., & Clark, J. M. (1988). An encoding-complex view of cognitive number processing: Comment on McCloskey, Sokol, and Goodman (1986). *Journal of Experimental Psychology: General, 117,* 204–214. http://dx.doi.org/10.1037/0096-3445.117.2.204

Campbell, J. I., & Epp, L. (2005). Architectures for arithmetic. In J. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 347–360). New York, NY: Psychology Press.

Cantlon, J. F., Brannon, E. M., Carter, E. J., & Pelphrey, K. A. (2006). Functional imaging of numerical processing in adults and 4-y-old children. *PLoS Biology, 4,* e125. http://dx.doi.org/10.1371/journal.pbio.0040125

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511571312

Chipman, S. F., Marshall, S. P., & Scott, P. A. (1991). Content effects on word problem performance: A possible source of test bias? *American Educational Research Journal, 28,* 897–915. http://dx.doi.org/10.3102/00028312028004897

Clark, J. M., & Campbell, J. I. (1991). Integrated versus modular theories of number skills and acalculia. *Brain and Cognition, 17,* 204–239. http://dx.doi.org/10.1016/0278-2626(91)90075-J

Coalition, S. T. E. M. (2000). *Before it's too late: A report to the nation from the National Commission on Mathematics and Science Teaching in the 21st century.* Retrieved from http://www2.ed.gov/inits/Math/glenn/report.pdf

Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research, 24,* 315–318. http://dx.doi.org/10.2307/3151642

Cote, J. A., & Buckley, M. R. (1988). Measurement error and theory testing in consumer research: An illustration of the importance of construct validation. *Journal of Consumer Research, 14,* 579–582. http://dx.doi.org/10.1086/209137

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory.* Mason, OH: Cengage Learning.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20,* 405–438. http://dx.doi.org/10.1016/0010-0285(88)90011-4

Davis-Dorsey, J., Ross, S. M., & Morrison, G. R. (1991). The role of rewording and context personalization in the solving of mathematical word problems. *Journal of Educational Psychology, 83,* 61–68. http://dx.doi.org/10.1037/0022-0663.83.1.61

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44,* 1–42. http://dx.doi.org/10.1016/0010-0277(92)90049-N

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition, 1,* 83–120.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20,* 487–506. http://dx.doi.org/10.1080/02643290244000239

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284,* 970–974. http://dx.doi.org/10.1126/science.284.5416.970

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283–299). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/11383-020

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8,* 430–457. http://dx.doi.org/10.1207/S15328007SEM0803_5

Engle, R. W., & Kane, M. J. (2003). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of Learning and Motivation, 44,* 145–199. http://dx.doi.org/10.1016/S0079-7421(03)44005-X

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 102–134). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139174909.007

Engle, R. W., & Oransky, N. (1999). The evolution from short-term to working memory: Multistore to dynamic models of temporary storage. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 515–555). Cambridge, MA: MIT Press.

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological Science, 13,* 150–156. http://dx.doi.org/10.1111/1467-9280.00427

Fuchs, L. S., Fuchs, D., Craddock, C., Hollenbeck, K. N., Hamlett, C. L., & Schatschneider, C. (2008). Effects of small-group tutoring with and without validated classroom instruction on at-risk students' math problem solving: Are two tiers of prevention better than one? *Journal of Educational Psychology, 100,* 491–509. http://dx.doi.org/10.1037/0022-0663.100.3.491

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1990). *Test of computational fluency. Available from L. S. Fuchs, Department of Special Education, 328 Peabody* (p. 37203). Nashville, TN: Vanderbilt University.

Fuchs, L. S., Hamlett, C. L., & Powell, S. R. (2003). *Grade 3 math battery.* Unpublished paper, Available from L. S. Fuchs, Department of Special Education, 328 Peabody, Vanderbilt University, Nashville, TN.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31,* 197–218. http://dx.doi.org/10.1207/s15327906mbr3102_3

Greer, B. (1997). Modelling reality in mathematics classrooms: The case of word problems. *Learning and Instruction, 7,* 293–307. http://dx.doi.org/10.1016/S0959-4752(97)00006-6

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE Publications.

Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computation skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology, 79,* 192–227. http://dx.doi.org/10.1006/jecp.2000.2586

Helwig, R., Rozek-Tedesco, M. A., & Tindal, G. (2002). An oral versus a standard administration of a large-scale mathematics test. *The Journal of Special Education, 36,* 39–47. http://dx.doi.org/10.1177/00224669020360010401

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research, 93,* 113–125. http://dx.doi.org/10.1080/00220679909597635

Jordan, N. C., & Hanich, L. B. (2000). Mathematical thinking in second-grade children with different forms of LD. *Journal of Learning Disabilities, 33,* 567–578. http://dx.doi.org/10.1177/002221940003300605

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review, 9,* 637–671. http://dx.doi.org/10.3758/BF03196323

Kelly, D., Xie, H., Nord, C. W., Jenkins, F., Chan, J. Y., & Kastberg, D. (2013). *Performance of U.S. 15-year-old students in mathematics, sci-*

ence, and reading literacy in an international context: First look at PISA 2012. (NCES 2014–024) (p. 23). Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014024

LeFevre, J. A., Berrigan, L., Vendetti, C., Kamawar, D., Bisanz, J., Skwarchuk, S. L., & Smith-Chant, B. L. (2013). The role of executive attention in the acquisition of mathematical skills for children in Grades 2 through 4. *Journal of Experimental Child Psychology, 114,* 243–261. http://dx.doi.org/10.1016/j.jecp.2012.10.005

Lourenco, S. F., Bonny, J. W., Fernandez, E. P., & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 18737–18742. http://dx.doi.org/10.1073/pnas.1207212109

Marsh, V., Beard, M., & Bailey, C. (2002). Multitrait-multimethod matrix in scientific inquiry. *Journal of Theory Construction & Testing, 6,* 94–97.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14,* 160–179. http://dx.doi.org/10.1080/10627190903422906

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4,* 169.

Mazzocco, M. M. M., & Kover, S. T. (2007). A longitudinal assessment of executive function skills and their association with math performance. *Child Neuropsychology, 13,* 18–45. http://dx.doi.org/10.1080/09297040600611346

McCloskey, M. (1992). Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition, 44,* 107–157. http://dx.doi.org/10.1016/0010-0277(92)90052-J

McCloskey, M., Caramazza, A., & Basili, A. (1985). Cognitive mechanisms in number processing and calculation: Evidence from dyscalculia. *Brain and Cognition, 4,* 171–196. http://dx.doi.org/10.1016/0278-2626(85)90069-7

McCloskey, M., Macaruso, P., & Whetstone, T. (1992). The functional architecture of numerical processing mechanisms: Defending the modular model. In J. Campbell (Ed.), *Advances in psychology (Vol. 91), The nature and origins of mathematical skills* (pp. 493–537). Amsterdam, the Netherlands: Elsevier.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.

Messick, S. (1996). Validity of performance assessment. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: National Center for Education Statistics (NCES 96–802).

Muth, K. D. (1984). Solving arithmetic word problems: Role of reading and computational skills. *Journal of Educational Psychology, 76,* 205–210. http://dx.doi.org/10.1037/0022-0663.76.2.205

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

National Academy of Sciences. (2001). *Adding it up: Helping children learn mathematics.* (J. Kilpatrick, J. Swafford, B. Findell, Mathematics Learning Study Committee, & National Research Council, Eds.). Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/catalog/9822.html

National Center for Education Statistics. (2013). *The nation's report card: A first look: 2013 mathematics and reading.* (NCES 2014–451) (pp. 1–12). Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved from papers3://publication/uuid/45C8F3B4-8A8D-459C-A3C3-2D59E6429AA4

National Research Council Committee on Appropriate Test Use. (1999). *High stakes: Testing for tracking, promotion, and graduation* (J. P. Heubert & R. M. Hauser, Eds.). Washington, DC: National Academy Press.

Nesher, P. (1986). Learning mathematics: A cognitive perspective. *American Psychologist, 41,* 1114–1122. http://dx.doi.org/10.1037/0003-066X.41.10.1114

Newcomer, P. L., & Hammill, D. D. (1988). *Test of language development* (Revised ed.). Austin, TX: Pro-Ed.

Passolunghi, M. C., Vercelloni, B., & Schadee, H. (2007). The precursors of mathematics learning: Working memory, phonological ability and numerical competence. *Cognitive Development, 22,* 165–184. http://dx.doi.org/10.1016/j.cogdev.2006.09.001

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron, 53,* 293–305. http://dx.doi.org/10.1016/j.neuron.2006.11.022

Pickering, S., & Gathercole, S. (2001). *Working memory test battery for children.* London, United Kingdom: Psychological Corporation.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879–903. http://dx.doi.org/10.1037/0021-9010.88.5.879

Reynolds, C. R., & Suzuki, L. A. (2012). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 67–94). New York, NY: Wiley. http://dx.doi.org/10.1002/9781118133880.hop210004

Rhodes, K. T., Branum-Martin, L., Morris, R. D., Romski, M., & Sevcik, R. A. (2015). Testing math or testing language? The construct validity of the KeyMath-Revised for children with mild intellectual disability and language difficulties. *American Journal on Intellectual and Developmental Disabilities, 120,* 542–568. http://dx.doi.org/10.1352/1944-7558-120.6.542

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). New York, NY: Academic Press.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11,* 105–126. http://dx.doi.org/10.1207/s15326977ea1102_2

Siegler, R. S. (1991). Strategy choice and strategy discovery. *Learning and Instruction, 1,* 89–102. http://dx.doi.org/10.1016/0959-4752(91)90020-9

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of cognitive skills: The eighteenth annual Carnegie symposium on cognition* (pp. 229–293). Hillsdale, NJ: Erlbaum.

Stanescu-Cosson, R., Pinel, P., van De Moortele, P. F., Le Bihan, D., Cohen, L., & Dehaene, S. (2000). Understanding dissociations in dyscalculia: A brain imaging study of the impact of number size on the cerebral networks for exact and approximate calculation. *Brain: A Journal of Neurology, 123,* 2240–2255. http://dx.doi.org/10.1093/brain/123.11.2240

Starkey, P., & Cooper, R. G., Jr. (1980). Perception of numbers by human infants. *Science, 210,* 1033–1035. http://dx.doi.org/10.1126/science.7434014

Stern, E., & Lehrndorfer, A. (1992). The role of situational context in solving word problems. *Cognitive Development, 7,* 259–268. http://dx.doi.org/10.1016/0885-2014(92)90014-I

Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., . . . Wigal, T. (2012). Categorical and dimensional definitions and evaluations of symptoms of ADHD: History of the SNAP and the SWAN Rating Scales. *The International Journal of Educational and Psychological Assessment, 10,* 51–70.

Terry, J. M., Hendrick, R., Evangelou, E., & Smith, R. L. (2010). Variable dialect switching among African American children: Inferences about working memory. *Lingua, 120,* 2463–2475. http://dx.doi.org/10.1016/j.lingua.2010.04.013

Verschaffel, L., De Corte, E., & Lasure, S. (1994). Realistic considerations in mathematical modeling of school arithmetic word problems. *Learning and Instruction, 4,* 273–294. http://dx.doi.org/10.1016/0959-4752(94)90002-7

Verschaffel, L., Greer, B., & de Corte, E. (2000). *Making sense of word problems.* Exton, PA: Swets & Zeitlinger.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence.* San Antonio, TX: Psychological Corporation.

Wilkinson, G. S. (1993). *Wide range achievement tests* (3rd ed.). Wilmington, DE: Jastak Associates.

Woodcock, R. W. (1997). *Woodcock diagnostic reading battery.* Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of cognitive abilities.* Itasca, IL: Riverside.

Woodward, J. (2004). Mathematics education in the United States: Past to present. *Journal of Learning Disabilities, 37,* 16–31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15493464; http://dx.doi.org/10.1177/00222194040370010301

Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 358,* 749–750.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74,* B1–B11. http://dx.doi.org/10.1016/S0010-0277(99)00066-9

Zelazo, P. D., & Frye, D. (1998). Cognitive complexity and control: The development of executive function in childhood. *Psychological Science, 7,* 121–126.

Zheng, X., Swanson, H. L., & Marcoulides, G. A. (2011). Working memory components as predictors of children's mathematical word problem solving. *Journal of Experimental Child Psychology, 110,* 481–498. http://dx.doi.org/10.1016/j.jecp.2011.06.001

Received March 11, 2016
Revision received November 27, 2016
Accepted December 16, 2016 ■

Call for Papers

Guest Editors
Mike C. Parent, PhD. Texas Tech University, Department of Psychological Sciences, Lubbock, Texas.
Francisco J. Sánchez, PhD. University of Missouri, Department of Educational, School, and Counseling Psychology. Columbia, Missouri.

Psychology of Men & Masculinity is soliciting papers for a Special Issue examining men and boys, masculinity, and physical health. Our goal with this special issue is to further our understanding of what contributes to masculine norms and how masculine norms affect men's and boys' physical health. Men's health issues are an important public health concern, and the interplay between the psychology of men and masculinity and men's physical health is complex. Research has already uncovered important links between the enactment of masculine norms and physical health. The enactment of masculinity is a vital component of men's health, and this Special Issue seeks to centralize the intersection of masculinity and health.

We are calling for contributions to this special issue that include quantitative and qualitative research encompassing social, psychological, medical, and public health perspectives. We especially encourage submissions that focus on the health experiences of minority individuals, broadly defined.

Examples of potential submission topics include:
1. Men and boys, masculinity, and cancer, including prostate, skin, and lung cancers
2. Men and boys, masculinity, and cardiovascular health and heart disease, including dietary and exercise perspectives
3. Masculinity in the context of disability and chronic disease conditions
4. Men and boys, masculinity, and obesity and diabetes
5. Men and boys, masculinity, and healthful aging
6. Men and boys, masculinity, and sexual health (e.g., use of PrEP)
7. Biological bases for men's and boys' health

The submission deadline is November 1, 2017. All submissions should adhere to APA 6th edition style requirements.

Please contact Dr. Mike Parent (michael.parent@ttu.edu) or Dr. Francisco Sanchez (sanchezf@missouri.edu) with any further questions.

Journal of Educational Psychology
2017, Vol. 109, No. 7, 977–992

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000182

A Meta-Analysis of the Relation Between RAN and Mathematics

Tuire Koponen
University of Jyväskylä and Niilo Mäki Institute, Jyväskylä, Finland

George Georgiou
University of Alberta

Paula Salmi
Niilo Mäki Institute, Jyväskylä, Finland

Markku Leskinen and Mikko Aro
University of Jyväskylä

Several studies have shown that rapid automatized naming (RAN) is a significant predictor of mathematics, but the nature of their relationship remains elusive. Thus, the purpose of this meta-analysis was to estimate the size of their relationship and determine the conditions under which they correlate. We used a random-effects model analysis of data from 38 studies (33 unique samples, 151 correlations, 7,135 participants) to examine the size of the RAN–mathematics relationship and the role of different moderators (i.e., math measure and variable, type of RAN task, math age, study design, and sample characteristics). The results showed a significant correlation between RAN and mathematics (r = .37; 95% confidence interval [CI] [.33–.42]) as well as a large heterogeneity of individual correlations. The results also revealed that RAN produced stronger correlations with arithmetic calculation tasks than with general achievement tests; stronger correlations with single-digit calculation tasks than multidigit calculation tasks; and stronger correlations with math fluency tasks than math accuracy tasks. The effect of these moderators suggests that part of the reason why RAN predicts mathematics is that they both require quick access to and retrieval of phonological representations from long-term memory. Our findings also suggest that RAN objects or colors can be used as early predictors of mathematical skill, especially of arithmetic fluency.

Keywords: rapid automatized naming (RAN), mathematics, arithmetic, meta-analysis

Rapid automatized naming (RAN), defined as the ability to rapidly name familiar visual stimuli such as letters, digits, colors, and objects, has been established as a strong concurrent and longitudinal predictor of reading in different languages (e.g., Compton, 2003; de Jong & van der Leij, 1999; Georgiou, Torppa, Manolitsis, Lyytinen, & Parrila, 2012; Juul, Poulsen, & Elbro, 2014; Landerl & Wimmer, 2008; Lervåg, Bråten, & Hulme, 2009; Liao, Georgiou, & Parrila, 2008; Parrila, Kirby, & McQuarrie, 2004; Savage & Frederickson, 2005), and a core deficit in dyslexia (e.g., de Jong & van der Leij, 2002; Eklund, Torppa, & Lyytinen, 2013; Kirby, Parrila, & Pfeiffer, 2003; Korhonen, 1995; Wimmer, Mayringer, & Landerl, 1998). Four independent meta-analyses have estimated the correlation between RAN and reading to be between .38 and .51 (see Araújo, Reis, Petersson, & Faísca, 2015; Scarborough, 1998; Song, Georgiou, Su, & Shu, 2016; Swanson, Trainin, Necoechea, & Hammill, 2003).

More recently, however, researchers have used RAN as a predictor of another important academic skill: mathematics. Despite the steady increase in the number of these studies, the findings are mixed and the conclusions indefinite. On the one hand, some studies have shown that RAN is an important predictor of mathematical skills and that the correlations between rapid naming and mathematics might be as high as those reported for reading (e.g., Berg, 2008; Koponen et al., 2016; Swanson, 2006b; Swanson, Jerman, & Zheng, 2008). Substantial correlations between RAN and mathematics have been reported in both cross-sectional (e.g., Cirino, 2011; Koponen, Aunola, Ahonen, & Nurmi, 2007) and longitudinal studies (e.g., Geary, 2011; Georgiou, Tziraki, Manolitsis, & Fella, 2013; Koponen et al., 2016). Such correlations have also been reported in various samples, such as typically developing children (e.g., Koponen et al., 2016; Niklas & Schneider, 2013), children at familial risk for dyslexia (e.g., van Bergen, de Jong, Maassen, & van der Leij, 2014; Koponen, Salmi, Eklund, & Aro, 2013), and children with mathematical difficulties (Mazzocco & Grimm, 2013). On the other hand, there are studies reporting either nonsignificant or weak correlations between RAN and mathematics (e.g., Niklas & Schneider, 2013) or high variance in the correlations (Hart, Petrill, Thompson, & Plomin, 2009). The contradictory findings might be related to the fact that math skill is multifactorial by nature (with several subskills) and thus RAN does not correlate equally well with all mathematical skills. It is also possible that there are some other moderating variables that influence the size of the relationship between RAN and mathe-

This article was published Online First March 13, 2017.

Tuire Koponen, Department of Education, University of Jyväskylä and Niilo Mäki Institute, Jyväskylä, Finland; George Georgiou, Department of Educational Psychology, University of Alberta; Paula Salmi, Niilo Mäki Institute; Markku Leskinen and Mikko Aro, Department of Education, University of Jyväskylä.

Correspondence concerning this article should be addressed to Tuire Koponen, Department of Education, University of Jyväskylä, PL 35, 40014 Finland. E-mail: tuire.k.koponen@jyu.fi

matics. In order to develop a more comprehensive picture of the relationship between RAN and mathematics, this meta-analysis examines the size of the relationship between RAN and different mathematics skills as well as the role of different moderators in the RAN–math relationship.

Examining the relationship between RAN and mathematics has important practical and theoretical implications. From a practical point of view, generating new information on the relationship between RAN and mathematics is important because it will enhance our understanding of the early predictors of mathematics development and the possible sources of difficulties in mathematics disabilities. If RAN proves to be a significant correlate of mathematics, then RAN tasks could be used as predictors of mathematics performance and early markers of future mathematical difficulties. Previous studies suggest that RAN, measured in children as young as 5 or 6 years old, can predict mathematics skill at school age (e.g., Georgiou et al., 2013; Koponen et al., 2013, 2016). Some of the previous studies suggest also that RAN makes a unique contribution to arithmetic fluency in Grades 2 and 3 above and beyond the contribution of verbal short-term memory (STM), working memory, and phonological awareness (Koponen et al., 2013, 2016). However, some researchers have questioned the role of RAN in mathematics (e.g., Georgiou et al., 2013; Willburger, Fussenegger, Moll, Wood, & Landerl, 2008). For example, Georgiou et al. (2013) found that processing speed (including numerical items) and visual memory explained most of RAN's predictive variance in calculation fluency. Due to the contradictory findings of previous studies, a meta-analysis could shed light on the RAN–mathematics relationship and on the role of different moderating variables.

From a theoretical point of view, examining the RAN–mathematics relationship allows us to test some interesting hypotheses regarding the nature of RAN and the underlying cognitive processes in mathematics. In reading research, scholars have argued that RAN is an index of the speed of access to and retrieval of phonological representations from long-term memory (e.g., Bowey, McGuigan, & Ruschena, 2005; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997). Empirical findings from behavioral studies have also suggested that besides reading, mathematics and arithmetical calculation also require quick retrieval of phonological representations from long-term memory (de Smedt, Taylor, Archibald, & Ansari, 2010), because arithmetic facts are also supposedly stored as phonological forms in long-term memory (de Smedt & Boets, 2010; Simmons & Singleton, 2008). In line with such findings, evidence from neuroimaging studies on RAN (e.g., Cummine, Szepesvari, Chouinard, Hanif, & Georgiou, 2014; Misra, Katzir, Wolf, & Poldrack, 2004) and mathematics (e.g., Dehaene, Piazza, Pinel, & Cohen, 2003; Wei et al., 2014) indicate that both skills are associated with regions of the left temporo-parietal cortex, such as the left angular gyrus. This region is activated during phonological decoding (e.g., Price & McCrory, 2005). Quick access to arithmetical facts is important because it facilitates the calculation process and enables the learner to release her working memory capacity for problem solving. Finding a predictor for arithmetical calculation fluency at school age could provide a possibility for early identification of risk for dysfluency difficulties and thus for early support.

However, given that mathematics consists of a wide set of different subskills and that retrieval of arithmetic facts from memory is not a core requirement in all mathematical tasks, RAN should not correlate equally well with all mathematical outcomes. In addition to the distinction between general mathematical achievement and arithmetic calculation, other aspects of the math outcome may also influence the RAN–mathematics relationship. As is the case with reading tasks (Araújo et al., 2015), RAN (which is a speeded measure) may correlate more strongly with the speed of mathematics performance, that is, tasks in which the score is either the response time or the number of items completed within a specified time limit. Moreover, the type of arithmetic problems might influence the association between RAN and mathematics. Answers to single-digit calculation problems (e.g., $3 + 5$; 5×2) are usually retrieved from memory. This is obviously not the case in multidigit calculation problems (e.g., $325 + 196$), which require knowledge of place value and mastery of arithmetic procedures in addition to retrieving/calculating partial answers. According to this proposal, if RAN is an index of the speed of lexical access (e.g., Norton & Wolf, 2012), it should correlate more strongly with single-digit calculation problems than with multidigit calculation problems that require several mental operations on top of retrieval, such as calculating partial answers and composing them in order to find the final answers. The type of RAN tasks may also influence the RAN–mathematics relationship. Araújo et al. (2015) and Scarborough (1998) showed that alphanumeric RAN (digits and letters) is more strongly related to reading than nonalphanumeric RAN (objects and colors). Once formal reading instruction begins, alphanumeric stimuli are explicitly taught and practiced, in contrast to the names of colors and objects that are learned more implicitly during language development. Obviously, the same argument holds for mathematics, and thus a stronger relationship would be expected between alphanumeric RAN and mathematics than between nonalphanumeric RAN and mathematics. However, there might also be some domain-specific features in the relationship between RAN and mathematics, and thus the findings would not necessary be the same for those found for reading. A recent study conducted by Donker, Kroesbergen, Slot, Van Viersen, and De Bree (2016) found that children with mathematical difficulties were impaired only in nonalphanumeric RAN, while children with reading or comorbid difficulties were impaired in both alphanumeric and nonalphanumeric RAN. Donker et al. (2016) suggested that nonalphanumeric RAN requires additional conceptual processing, as opposed to alphanumeric RAN, which requires more phonological processing. Furthermore, children with mathematical disabilities may have difficulty with the conceptual processing of quantities represented by the digits, but not with access to number words per se.

This proposal is closely related to the discussion of the underlying deficit in mathematical learning difficulties. According to the access deficit hypothesis, the origin of mathematical disability lies in problems accessing magnitude representations from symbolic information (numbers; Rousselle & Noël, 2007). If this is true, then RAN should be more strongly related to mathematics when RAN stimuli are numbers rather than letters, colors, or objects. However, another approach suggests that the core deficit of mathematical disabilities is in magnitude processing, which can be seen both in nonsymbolic and symbolic magnitude processing (Butterworth, 2005. According to this view, an equally important division of the RAN tasks could be grouping them into numeric (digits, dice) and non-numeric (letters, colors, and objects) tasks. The

assumption that the RAN–mathematics relationship is restricted to the use of a numeric stimulus in RAN tasks has received support in previous studies (Landerl, Fussenegger, Moll, & Willburger, 2009; Willburger et al., 2008). However, it is also possible that the association between RAN and mathematics might not be specific to numbers, but could instead reflect an inherent ability to learn and retrieve arbitrary visual—verbal associations (Manis, Seidenberg, & Doi, 1999). If this is true, then equally strong correlations should be observed between numeric (quantities, digits) and nonnumeric RAN (letters, colors, and objects) tasks with mathematics.

A third moderator of the RAN–mathematics relationship may be the age when mathematics was first assessed. Between Grades 2 and 3, arithmetic calculation skills generally progress from effortful counting strategies to more automatic retrieval strategies (Jordan, 2003). Thus, in age-appropriate development of arithmetic skills, children usually start using fact retrieval as their main strategy for solving mathematical problems between the ages of 9 and 10 (Lemair & Siegler, 1995). Consequently, after the age of 9, when arithmetic calculation skill has reached an automatic level, RAN should be more strongly related to arithmetic, compared to the developmental phase when counting-based strategies (e.g., the counting-on strategy) are more common.

Finally, factors such as study design and participant characteristics should be taken into account when examining the RAN–mathematics relation. In general, correlations between two measures obtained at the same measurement point (as in concurrent studies) are stronger than correlations between two measures assessed at different time points (as in longitudinal studies). Previous studies in reading have shown that RAN is more strongly related to reading among low-performing children (e.g., McBride-Chang & Manis, 1996; Savage & Frederickson, 2005; Scarborough, 1998). To our knowledge, no previous studies have examined whether the RAN–mathematics relationship is stronger among children with math disabilities. In light of the findings of previous reading studies, it would be interesting to assess whether RAN is also more strongly associated with math among low-performing children. We could assume a higher prevalence of retrieval difficulties among atypical samples and overlapping dysfluency problems in naming and calculation.

The Present Study

The current study aims to answer the following six research questions:

Research question 1. To what extent is RAN related to mathematics? Because of the contradictory findings of previous studies, we did not formulate a specific hypothesis.

Research question 2. Is the RAN–mathematics relationship affected by the nature of the mathematics measure (achievement tests vs. arithmetic; multidigit vs. single-digit calculations) or task requirements (fluency vs. accuracy)? Because fluent calculation relies on rapid retrieval of the answers (arithmetic facts) from long-term memory, we hypothesized that RAN tasks would be more strongly related to arithmetic calculation than to general math achievement; more strongly related to single-digit calculation than to multidigit calculation; and more strongly related to math fluency than to math accuracy.

Research question 3. Is the RAN–mathematics relationship affected by the type of RAN task? More specifically, does numeric RAN (numbers and/or quantities) correlate more strongly with mathematics than non-numeric RAN (colors, objects, letters)? Furthermore, does alphanumeric RAN (digits and letters) correlate more strongly with mathematics than nonalphanumeric RAN? Based on the findings of previous studies that showed children with dyscalculia have a specific deficit in RAN quantities (e.g., Landerl et al., 2009; Willburger et al., 2008), we hypothesized that numeric RAN would be more strongly related to math than non-numeric RAN. Because the findings of previous studies on the role of alphanumeric and nonalphanumeric RAN in mathematics are mixed, we did not formulate a specific hypothesis.

Research question 4. Is the RAN–mathematics relationship affected by the age when math tasks are assessed? We hypothesized that RAN would be more strongly related to mathematics after the age of 9 (around the time when arithmetic calculation skill becomes automatic) than in younger ages.

Research question 5. Is the RAN–mathematics relationship affected by the study design (cross-sectional vs. longitudinal)? Given that correlations tend to be higher in skills assessed concurrently than when there is a time distance between measurements, we hypothesized that RAN would correlate more strongly with mathematics in concurrent studies than in longitudinal studies.

Research question 6. Is the RAN–mathematics relationship affected by the sample characteristics (a sample including high prevalence of children with learning disabilities or low-performing children vs. a normal sample)? As RAN is more strongly related to reading among poor or at-risk readers (e.g., Meyer et al., 1998; Savage & Frederickson, 2005; Scarborough, 1998), we hypothesized that RAN would be more strongly associated with math among low-performing children. This is based on the assumption that dysfluent calculation, which often means retrieval difficulties, is at least partly related to co-occurring naming difficulties. However, a strong hypothesis cannot be proposed because the studies with atypical samples consist of quite heterogeneous populations with different kinds of difficulties and not only children with mathematical difficulties. Moreover, previous math literature does not provide information regarding whether the relationship between RAN and math would vary among poor-, average-, and well-performing children.

Method

Data Collection

The inclusion, search, and coding procedures are detailed in Figure 1. For the target constructs examined in this study (RAN and mathematics), we established the operational criteria to determine the indicators of each construct. A task was considered a measure of RAN if quick serial naming of an array of objects, colors, letters, digits, or quantities was required. In turn, to be considered a measure of math achievement, the test should require mathematical skills other than just arithmetic calculation (e.g., knowledge of the number system, fractions, or geometry). Arithmetic calculation included only tasks that required solving different arithmetic operations (addition, subtraction, multiplication, and division). Arithmetic calculation was further divided into single-digit (e.g., $3 + 4$; $6 - 2$; 5×4) and multidigit calculation ($32 + 41$, $76-12$, 14×24) tasks. Mathematical accuracy included mea-

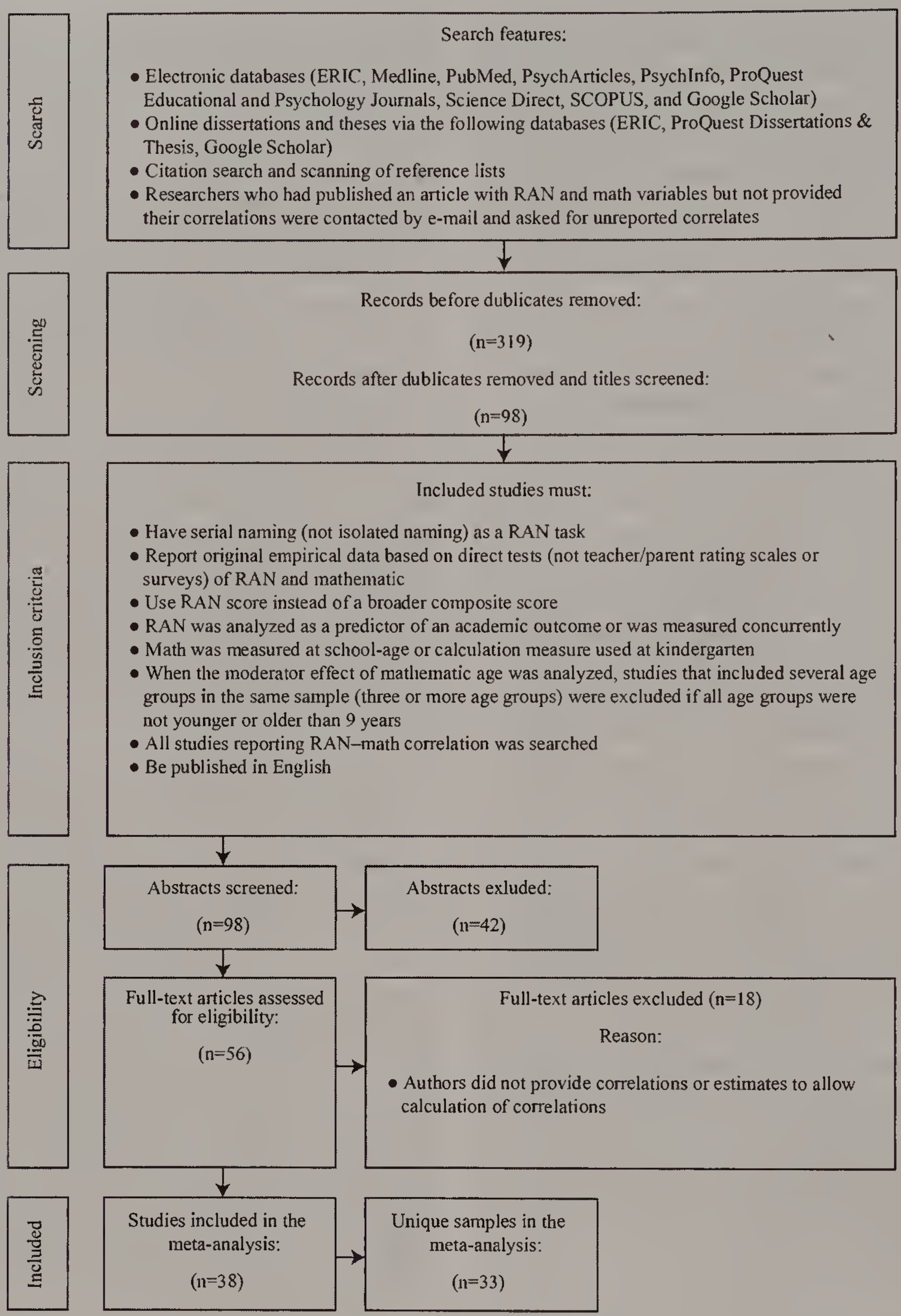


Figure 1. Flow diagram for the search for and inclusion of studies.

sures based on the accuracy of mathematical problem solving or calculation. To be considered a measure of math fluency, the task should require children to solve as many arithmetic or other math problems as possible within a specified time limit.

Inclusionary Criteria and Screening Process

The search followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement protocol (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009), and three methods were used to identify relevant studies. First, we searched electronic databases and e-journal services (ERIC, Medline, PubMed, PsychArticles, PsychInfo, ProQuest Educational and Psychology Journals, Science Direct, Scopus, and Google Scholar) for publications in English by using the keywords *rapid serial naming**, *naming speed**, *rapid automatized naming**, and *RAN** combined with *mathematics** and *arithmetic** in the subject or title. Second, we searched online dissertations and theses via databases (ERIC, ProQuest Dissertations & Theses Global, and Google Scholar) with the same keywords. No restrictions were imposed regarding the publication year. The search covered studies published before July 2016. Third, we checked the reference lists of the collected reports for relevant studies. We contacted authors who had published an article with RAN and math measures but had not provided correlations via e-mail and kindly asked them to send us the correlations.

In addition, we used the following eight inclusionary criteria:

(a) the RAN tasks required serial naming instead of isolated naming;

(b) Original empirical data were based on direct assessment (not teacher/parent rating scales or surveys);

(c) Correlations were reported at the level of naming subtasks (studies including RAN composite scores consisting of several subskills were excluded);

(d) RAN was used as a predictor of a mathematics outcome or both constructs (RAN and mathematics) were measured concurrently (studies in which mathematics performance was assessed prior to RAN were excluded);

(e) Math was measured at school age or the calculation was measured at kindergarten (early number skills were not included);

(f) When analyzing the moderator effect of mathematics age, studies reporting RAN–math correlations among samples consisting of several age groups in the same sample (three or more age groups) were excluded if all age groups were not younger or older than 9 years;

(g) When multiple measures were used to assess one construct, all qualified correlations were included in the dataset. In the analyses, dependencies between correlations from a single study or a single data set were taken into account; and

(h) All studies that provided effect sizes (*d*) and regression coefficients (R^2, beta) as an estimate of the RAN–math relationship were also included after they had been converted into the Pearson product–moment correlation coefficient effect sizes *r* (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cohen, 1988).

After these criteria were used, 33 unique samples with 151 outcomes were found (min = 1, max = 40) with sample sizes ranging from 29 to 628 (see Table 1).

Coding Procedures

After the qualified studies had been selected, we coded relevant information from the studies. This information included the following variables of descriptive data: (a) number of participants; (b) mean ages of the participants at the time RAN and math were measured; (c) design: longitudinal versus cross-sectional; (d) sample characteristics: typical or atypical; (e) the type of measure used for each construct and the outcome variable (e.g., accuracy or fluency); and (f) the year of publication. In the published studies, some measures were coded so that a negative correlation indicated a positive relationship between RAN and mathematical performance. In these cases, the effect size measures were recoded such that a positive correlation always indicated a positive relationship. The first author, together with an expert in math research, double coded the moderator variables. The agreement rate between the coders varied between 89% and 94%. Differences between the coders were mostly due to limited information provided in the studies regarding the sample characteristics and tasks. Differences in the scoring by the raters were resolved after discussions with the first author.

Moderators

We coded two types of moderators: procedural and sample characteristics (see Table 2). Procedural moderators included the math domain assessed (general math achievement vs. arithmetic and single-digit vs. multidigit calculation), outcome variable in arithmetic (fluency vs. accuracy), RAN stimulus (non-numeric vs. numeric and nonalphanumeric vs. alphanumeric), and study design (cross-sectional vs. longitudinal). Sample moderators included the mathematical age of the children as a continuous variable and the sample type (two categories: an atypical sample that consisted of a high prevalence of low achievers or children with learning disabilities vs. a typical sample, i.e., a population-based sample or a sample with high prevalence of high achievers).

Effect Size Calculations

Effect sizes (*d*) and regression coefficients (R^2, beta) were first transformed into the Pearson product–moment correlation coefficient effect sizes *r* (Borenstein et al., 2009; Cohen, 1988). Next, the *r* effect sizes were transformed into Fisher's *z* values to be used in a meta-analysis, and the variance was calculated with Cox's (2008) formula $1/(n - 3)$.

Meta-Analytic Integration

Summative results and graphics (funnel plot, symmetry tests, violin plots, and forest plot) were produced in R (R Foundation, 2015). The metafor (Viechtbauer, 2010) package was used for funnel plot symmetry tests and forest plot construction. The Fisher *z* values were averaged by the studies individually and then transformed back into Pearson's correlation coefficients. The estimation method was the random-effects model with the restricted maximum likelihood (REML) method. The funnel plot asymmetry was tested with a trim fill test and a regression test in which the predictor was the standard error (Sterne & Egger, 2005; Sterne et al., 2011). We used the vioplot package (Adler, 2005) for the violin plot production (Hintze & Nelson, 1998).

A random-effects meta-analysis and a metaregression analysis for the moderators were performed in R (R Foundation, 2015). We used the robumeta application package (Fisher & Tipton, 2015) and the robust variance estimation (RVE) method (Hedges, Tipton, & Johnson, 2010; Tanner-Smith & Tipton, 2014) because there were several correlations from a single study and thus the data points were not independent. We adopted a dependent effects metaregression (D-MR) approach. The correlation coefficients were clustered into unique samples; that is, the number of studies represented the number of unique study samples. The weight of each study was the square of the standard error (Lipsey & Wilson, 2001). The heterogeneity was estimated from τ^2 and I^2 statistics (Higgins & Thompson, 2002). τ^2 was interpreted as between-study variance in study-average effect sizes. I^2 assessed the percentage of the total variance attributable to true heterogeneity. The efficient weights were analyzed with a sensitivity approach (effects of various ρ values (0–1) on the results). Because the results were not sensitive to the ρ values used, the results were reported for $\rho = .80$. Finally, small sample adjustments were used in the analysis (Tipton, 2015). The moderator effects were analyzed as within-study and between-study partition effects (i.e., the partition was done by

Table 1
Description of the Studies Presenting RAN-Math Correlations

Author (Year)	*n*	*r* math	Sample type	Design	Age	Math measure	Math variable	RAN stimulus
Ackerman (2001)	101	.43	LD and controls	Cross-sectional	116.4	Multi-digit calc.	A	L + D + D & L
Berg (2008)	90	.44	Normal	Cross-sectional	121.5	Multi-digit calc.	A	D
van Bergen et al. (2014)	196	.48	LD and controls	Longitudinal	107.1	Multi-digit calc.	F	C
Chan and Ho (2010)	168	.43	LD and controls	Cross-sectional	110.4	Multi-digit calc.	A	D
						Single-digit calc.	F	
Cirino (2011)	285	.41–.47	Normal	Cross-sectional	73.4	Single-digit calc.	F	D
								L
								O
Foster, Jason, Clements, and Sarama (2015)	208	.31–.34	Normal	Longitudinal	67.4–74.4	Math achievement	A	O
Fuchs et al. (2005)	272	.24–.29	LA and controls	Cross-sectional		Math achievement	F	D
						Multi-digit calc.	A	
						Single-digit calc.	F	
de Jong and van der Leij (1999)	166	.28–.31	Normal	Longitudinal	100.4	Single-digit calc.	F	O
Geary (2011)	177	.16–.47	Normal	Longitudinal	74.0–134.0	Math achievement	A	D
								L
Georgiou, Tziraki, Manolitsis, and Fella (2013)	72	.40	Normal	Longitudinal	83.1	Multi-digit calc.	F	O + C
Hannula, Lepola, and Lehtinen (2010)	139	.25	Normal	Longitudinal	102	Multi-digit calc.	A	O + C
Hart et al. (2009)	628	.11–.43	Twin	Longitudinal	102.5–118.7	Multi-digit calc.	A	L + D
						Single-digit calc.	F	
Kleemans, Segers, and Verhoeven (2012)	160	.17–.39	LD and controls	Longitudinal	85.1	Single-digit calc.	F	O
Koponen et al. (2006)	29	.64	LD	Cross-sectional	123.6	Single-digit calc.	F	O + C
Koponen et al. (2007)	207	.13–.37	Normal	Longitudinal	129	Single-digit calc.	F	O
						Multi-digit calc.	A	
Koponen et al. (2013)	362	.28–.42	LD and controls	Longitudinal	108.0–118.0	Multi-digit calc.	F	C
Koponen et al. (2016)	378	.27–.36	Normal	Longitudinal	115.0–127.0	Multi-digit calc.	F	O
								O
Krajewski and Schneider (2009)	130	.33–.44	Normal	Longitudinal	91.0–103.0	Math achievement	A	Dots & D
Landerl and Willburger (2010)	439	.32	LD and controls	Cross-sectional	111	Multi-digit calc.	F	L + D + O
Lepola, Niemi, Kuikka, and Hannula (2005)	139	.15–.25	Normal	Longitudinal	104	Multi-digit calc.	A	O + C
						Single-digit calc.	A	
Moll, Göbel, and Snowling (2015)	89	.26–.49	LD and controls	Cross-sectional	NA	Single-digit calc.	F	L
						Multi-digit calc.	A	D
Niklas and Schneider (2013, 2014)	608	.07–.27	Normal	Cross-sectional	77.0–87.0	Math achievement	A	O
Swanson and Beebe-Frankenberger (2004)	353	.63	LA and controls	Cross-sectional	107.9	Multi-digit calc.	A & F	L + D
Swanson (2006a)	127	.50–.59	HA and normal	Cross-sectional	88.4	Multi-digit calc.	A	L
							F	D
Swanson (2006b)	320	.52	LA and controls	Longitudinal	120.1	Multi-digit calc.	A & F	L + D
Swanson and Kim (2007)	353	.38–.64	Normal	Cross-sectional	NA	Multi-digit calc.	A	D
						Multi-digit calc.	F	L
Swanson (2008)	205	.64–.65	Normal	Cross-sectional	91.8	Multi-digit calc.	F	D
								L
Swanson et al. (2008)	353	.43–.60	LA and controls	Longitudinal	132.7	Multi-digit calc. (verbal)	F	L + D
Träff (2013)	134	.37–.48	LA and controls	Cross-sectional	142	Single-digit calc.	F	C
							A	

Table 1 (*continued*)

Author (Year)	*n*	*r* math	Sample type	Design	Age	Math measure	Math variable	RAN stimulus
van Daal, van der Leij, and Adèr (2013)	82	.46–.61	LD and controls	Cross-sectional	167	Single-digit calc.	F	O
								C
								D
								L
								L + D
van der Sluis, de Jong, P. F., and van der Leij (2007)	127	.38–.45	Normal	Cross-sectional	128	Multi-digit calc.	F	Q
								D
								L
								C
Waber et al. (2000)	188	.16–.33	LD	Cross-sectional	114	Math achievement	A	D
								L
								C
								O
Wocadlo and Rieger (2007)	63	.35	At- risk group	Cross-sectional	97	Math achievement	A	L + D

Note. HA = high achieving; LA = Low achieving; LD = Learning difficulties; A = Accuracy; F = Fluency; Co = Comprehension; D = Digits; C = Colors, O = Objects; L = Letters; Q = quantities.

using group.center and group.mean robumeta functions). Partitioned moderators were entered into the analysis separately in order to maximize the number of studies and correlation coefficients in the calculations.

The funnel plot and forest plot graphics and the overall testing of asymmetry were kept at the level of unique samples; that is, the data was the sample average of coefficients. The violin plot was used for the illustrations of distribution of all correlation coefficients. The average RAN–mathematics result presented in the forest plot (metafor) is the same as the average RAN–mathematics result in Table 2 (robumeta), but they were obtained via different random effects meta-analytic approaches.

Results

The literature search yielded 306 reports. We then narrowed down the literature to 92 potentially relevant reports (after duplicates were removed and the titles screened). Further screening of the abstracts resulted in 53 candidate reports. After the full texts were read, 38 reports with 33 unique samples were included in the meta-analysis. Figure 2 provides a funnel plot graph of the unique samples in which the correlation coefficients were averaged within the unique samples. The regression test for funnel plot asymmetry showed that the standard error was not a statistically significant predictor ($z = 1.67$, $p = .095$). The trim and fill analysis revealed that there was an estimated number of five missing studies on the left side of the funnel plot. The probability that there were no missing studies on the left side of the funnel plot was statistically significant ($p = .016$). The unique sample-averaged coefficients (Fisher's z) varied from .17 to .77, as is observable from the random-effect model forest plot graph presented in Figure 3. In sum, there was large variation among the correlation coefficients and substantive heterogeneity ($Q(32) = 136.51$, $p < .001$).

The summarized results are presented in Table 2 (Question 1). The mean random-effects model RAN–math weighted Fisher's z coefficient equaled .38 ($p < .001$, 95% CI [.33, .43), and after being back-transformed to Pearson r, equaled .37 (95% CI [.32, .41]). The violin plot of Fisher's z coefficients presented in Figure 4 indicates that the distribution is clearly positively skewed. (see Appendix Figures A1 and A2)

Next, we examined the role of the different moderators. In the within-study and between-study metaregression analyses, variations in the moderator effects were taken into account. The results showed that after the between-study partition estimation, the achievement versus arithmetic task (i.e., domain of math assessed 1) was a significant moderator ($\beta = .12$, $p < .05$). Arithmetic math measurements were associated with higher correlation coefficients. At the within-study partition level, the type of calculation performed (single-digit vs. multidigit calculations) was a significant moderator ($\beta = .14$, $p < .01$). The correlations were higher for the single-digit calculations. The math outcome measure (fluency vs. accuracy) was also a significant moderator associated with the between-study variation ($\beta = -.13$, $p < .05$). Fluency math outcomes produced higher correlations than math accuracy outcomes. Numeric RAN tasks also produced significantly larger correlations with math outcomes than non-numeric RAN ($\beta = .04$, $p < .01$); however, the statistically significant p (despite the low beta) might be an artifact (Type I error) associated with the estimation methods (see Tipton, 2015). The type of RAN stimulus (alphanumeric vs.

Table 2
Number of Correlations, Meta-Regression Estimate Based on Fisher's z, Standard Error, T-Test Results, 95% Confidence Interval (CI), Heterogeneity Statistics of the Relationship Between RAN-Math and Procedural and Sample Moderators

	$n(k)$	Estimate	Std Err	t-value	df	p	95% CI	τ^2	I^2
Null model									
Intercept	33 (151)	.38	.03	15.19	31.25	<.001	[.33, .43]	.02	81.25
Procedural moderators									
Domain of math assessed 1									
Center	33 (148)	.01	.05	.28	1.00[2]	.828	[−.66, .69]	.02	81.29
Mean	34 (151)	.12	.05	2.51	8.41	.035	[.04, .23]	.02	78.99
Domain of math assessed 2									
Center	27 (92)	.14	.03	5.11	5.87	.002	[.08, .21]	.02	79.02
Mean	27 (98)	−.05	.07	−.74	17.01	.469	[−.19, .09]	.02	80.32
Math outcome									
Center	33 (146)	−.07	.05	−1.40	8.84	.194	[−.19, .04]	.02	80.30
Mean	33 (151)	−.13	.05	−2.34	21.68	.029	[−.24, −.01]	.02	79.64
RAN stimulus 1									
Center	27 (115)	.04	.01	3.59	6.79	.009	[.01, .07]	.02	82.19
Mean	27 (124)	.14	.08	1.78	8.36	.111	[−.04, .32]	.02	81.48
RAN stimulus 2									
Center	31 (145)	.01	.07	.10	3.96[2]	.924	[−.19, .21]	.02	82.18
Mean	31 (146)	.09	.06	1.51	23.97	.144	[−.03, .21]	.02	81.13
Design									
Center[1]	—	—	—	—	—	—	—	—	—
Mean	34 (152)	−.08	.04	−1.83	26.16	.079	[−.17, .01]	.02	81.20
Sample moderators									
Age of math assessed									
Center	30 (136)	−.00	.04	−.05	2.92[2]	.963	[−.14, .13]	.02	79.30
Mean	30 (136)	.04	.05	.81	22.29	.428	[−.07, .16]	.02	79.87
Type of sample									
Center	33 (151)	.02	.12	.14	1.00[2]	.913	[−1.5, 1.53]	.02	81.28
Mean	33 (151)	.02	.05	.45	28.63	.656	[−.08, .12]	.02	81.68

Note. $n(k)$: number of samples (number of coefficients); Estimate: Robumeta correlated data meta-regression estimate, small-sample adjustment, ρ=.80; Std Err: standard error; CI: confidence interval; Domain of math assessed 1: 0 = achievement, 1 = arithmetic. Domain of math assessed 2: 0 = multi-digit, 1 = single digit. Math outcome: 0 = fluency, 1 = accuracy. RAN stimulus 1: 0 = nonnumeric, 1 = numeric. RAN stimulus 2: 0 = non-alphanumeric, 1 = alphanumeric. Design: 0 = cross-sectional, 1 = longitudinal. Age of math assessed: 0 = at or below 9 years, 1 = above 9 years. Type of sample: 0 = typical sample, 1 = atypical sample. Center: within-study centered estimation. Mean: between-study centered estimation.
[1] Estimation was not possible. [2] The result should be viewed with some caution due to low degrees of freedom value; $df < 4$.

nonalphanumeric), the design of the study (cross-sectional vs. longitudinal), age when math was assessed (at or below 9 years vs. above 9 years), and type of sample (population-based sample or sample with high prevalence of high achievers vs. high prevalence of low achievers or children with learning disabilities) were not significant moderators.

Discussion

The number of studies using RAN as a predictor of mathematics performance has steadily increased over the last decade (e.g., Berg, 2008; Cirino, 2011; Geary, 2011; Georgiou et al., 2013; Hecht, Torgesen, Wagner, & Rashotte, 2001; Koponen et al., 2007, 2013; Swanson et al., 2008). However, the nature of the RAN–mathematics relationship remains elusive. The current meta-analysis aimed first to examine the size of the RAN–mathematics relationship. We found a positive and significant relationship (r = .37) between RAN and mathematics and the effect size was large (Cohen, 1988). Interestingly, the average size of the RAN–mathematics relationship is close to that reported in previous meta-analyses of RAN and reading (Araújo et al., 2015; Scarborough, 1998; Song et al., 2016; Swanson et al., 2003).

However, the size of the RAN–mathematics relationship appears to be influenced by different moderators. First, the math domain used in previous studies was related to the strength of the RAN–mathematics relationship. As expected, RAN was a stronger correlate of math when math was operationalized with an arithmetic calculation task than when math was operationalized with general math achievement tests. More specifically, arithmetic tasks (e.g., Woodcock-Johnson math fluency) require responding to simple addition, subtraction, and multiplication problems or finding partial and total answers in multidigit calculations. Retrieving the names of numbers, operation symbols (e.g., +, −), and answers from long-term memory are central processes in arithmetical calculation, which tap the same capacities as RAN. In contrast, math achievement tests (e.g., the Wechsler Individual Achievement Test [WIAT] numerical operations) involve a wider set of mathematical subskills, including problems that cannot be simply solved with the retrieval of an answer from long-term memory. Consequently, compared to RAN, these tasks require a number of different processes, which results in lower correlations between them.

In previous arithmetic calculation studies, single-digit calculations produced significantly stronger correlations than multidigit

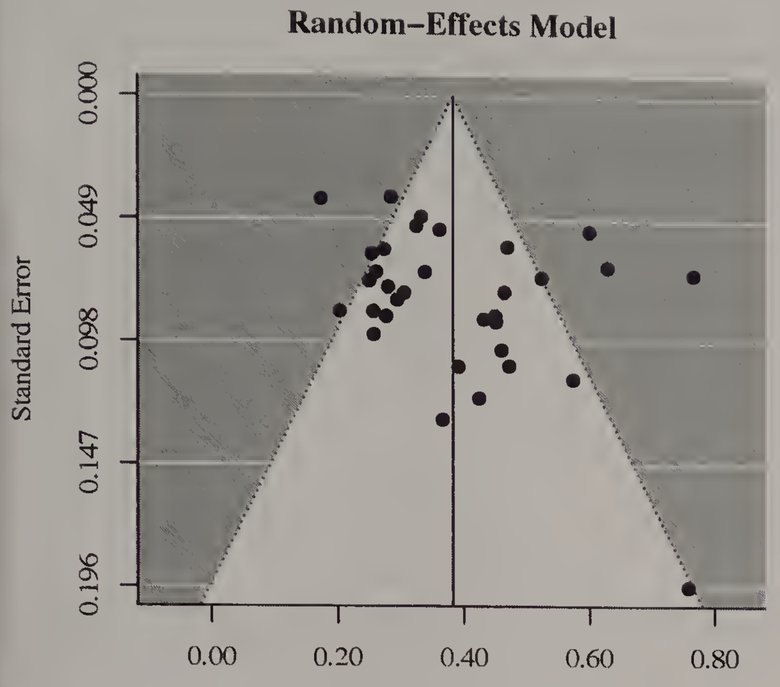


Figure 2. A funnel plot graph of the averaged and Fisher's z transformed RAN–Math correlation coefficients in the unique samples (N = 33).

calculations. This finding is in line with those of previous studies conducted by Koponen and colleagues (Koponen et al., 2007, 2013, 2016), which showed that RAN was related with reading and single-digit calculation fluency, but not with multidigit calculation (Koponen et al., 2007). Although single-digit calculation is needed to obtain intermediate answers in multidigit calculation, understanding place value, the ability to retrieve procedural knowledge, the use of algorithms, and monitoring multistep processes are required in multidigit calculation. In other words, multidigit calculation includes retrieving factual, procedural, and conceptual knowledge, whereas RAN and fluent single-digit calculation rely more on retrieving factual knowledge, such as names of numbers or objects.

The type of math outcome also explained a significant amount of variance in the RAN–mathematics relationship. Fluency outcomes produced stronger correlations with RAN than accuracy outcomes. This is similar to the finding reported for the RAN–reading relationship (Araújo et al., 2015; Song et al., 2016). These moderator effects support the view that RAN reflects, to some extent, the efficiency of access to and retrieval of phonological representations from long-term memory, which are needed in fluent calculation (de Smedt & Boets, 2010; Koponen et al., 2013).

The distinction between alphanumeric and nonalphanumeric RAN tasks failed to explain the significant variance in the RAN–mathematics relationship. This might partly be due to the different types of processes involved in RAN and in basic reading and arithmetic skills. More specifically, some researchers have argued that RAN letters and digits activate the same neural networks that are involved in phonological and orthographic processing in reading (e.g., Cummine et al., 2014; Cummine, Szepesvari, Chouinard, & Georgiou, 2015; Misra et al., 2004). In contrast, RAN objects also engage networks that are involved in semantic processing (e.g., Cummine et al., 2014; Humphreys, Price, & Riddoch, 1999). Since reading is usually assessed with word recognition and decoding tasks and not with comprehension tasks that would require semantic processing, alphanumeric RAN proves to be more reliable in predicting reading over nonalphanumeric RAN. In math, many tasks involve access to magnitude, and the use of magnitude information is beneficial in solving mathematical problems. Magnitude processing is also helpful in arithmetic calculation. For example, knowing that 6 is 1 more than 5 and being able to retrieve the answer for 5 + 5, which equals 10, can help an individual derive the answer for 5 + 6. Thus, in mathematics, parallel access to phonological and semantic information may boost the association between nonalphanumeric RAN and mathematics and explain why an equally strong correlation was found between nonalphanumeric and alphanumeric RAN tasks with math in the present meta-analysis. This suggestion is in line with the findings of a recent study conducted by Donker et al. (2016), in which nonalphanumeric RAN correlated with mathematics but alphanumeric RAN did not in a sample of children with math learning disabilities. Donker et al. suggested that nonalphanumeric RAN requires additional conceptual processing compared to the mere phonological processing that is required in alphanumeric RAN. Children with difficulties in math may have difficulties in conceptual processing over and above the difficulties in accessing and retrieving the digit names.

Using numeric stimuli (numbers or quantities) versus nonnumeric stimuli (letters, objects, or colors) had a very small effect on the RAN–math relationship and explained only the within-studies (not between) variation. This suggests that the RAN–math relationship cannot be explained by the use of numeric stimuli alone, but is related to the naming process itself. Thus, the findings from the present meta-analysis are only partially in line with the findings of Willburger et al. (2008) and Landerl et al. (2009), who showed that children with dyscalculia exhibited a unique deficit in the rapid naming of quantities, whereas a more general deficit in the rapid naming of objects, letters, digits, and quantities was evident in the dyslexia group and the comorbid dyslexia/dyscalculia group.

Math age or sample characteristics (high prevalence of children with difficulties or low-achieving children vs. a normal sample) did not account for the large variation among the correlations. The nonsignificant moderator effects of age were unexpected because fact retrieval ability underlies the RAN–mathematics relation, and in math, fact retrieval is less automatized in younger children. There can be several reasons for this finding, one of which is the confounding effect within moderators. More studies and correlations are needed in order to examine the interactions between the moderators. The nonsignificant moderator effect of sample type (typical vs. atypical) was not expected either. A possible explanation may be that the "atypical sample" grouping covered quite heterogeneous populations with different kinds of difficulties and not solely children with mathematical or language-based difficulties. Along these lines, a theoretically important but unexplored question is whether the RAN–mathematics relationship is similar or different among poor- and well-performing children. The findings of the present study did not provide any evidence that there are differences among these groups. However, as stated above, in

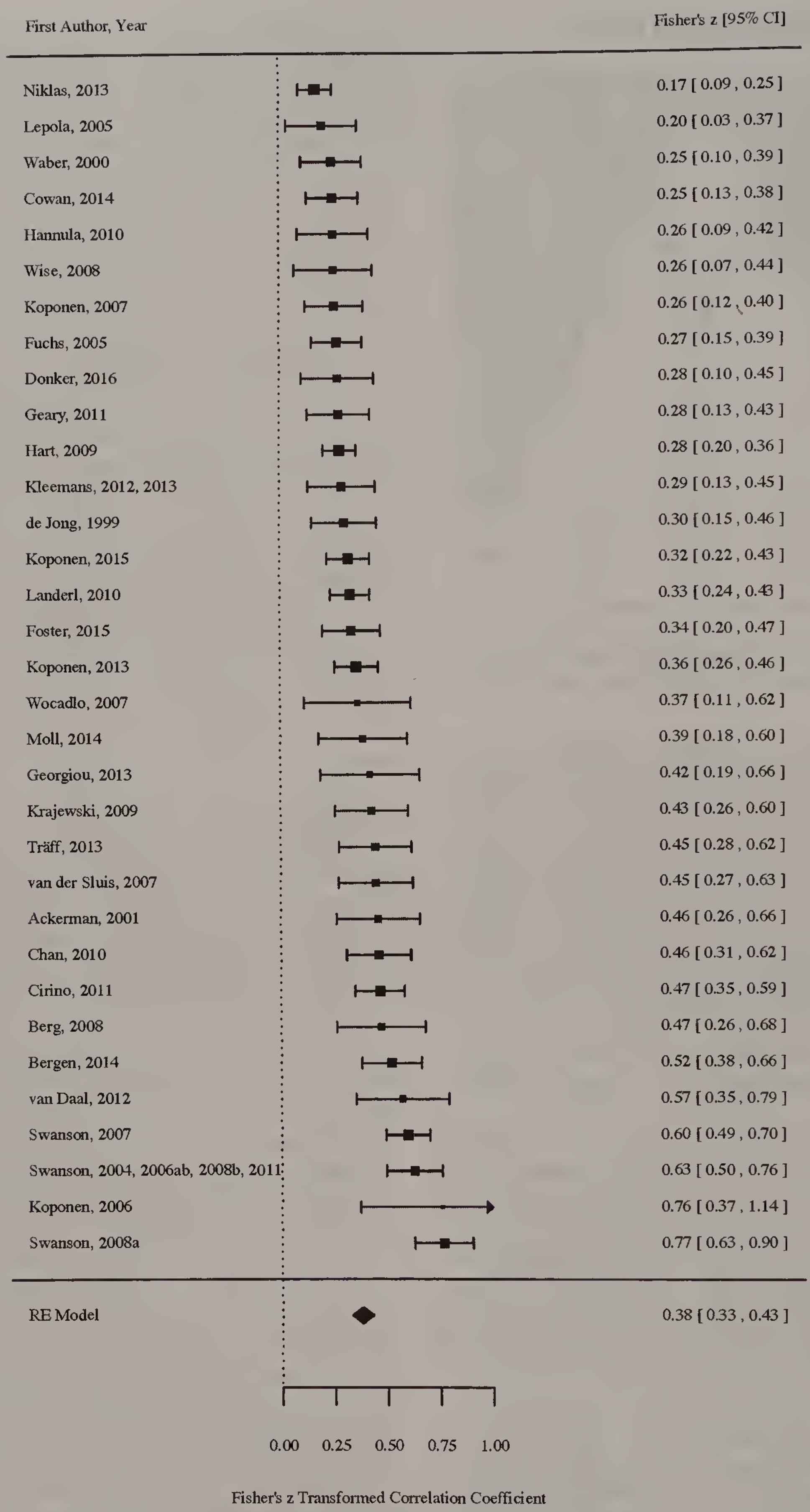


Figure 3. Overall random-effect model meta-analysis average of Fisher's z transformed correlation coefficient of RAN–Math correlation coefficients in the unique samples (N = 33) and coefficient with 95% confidence interval for each study. Coefficients represent average values of each unique sample.

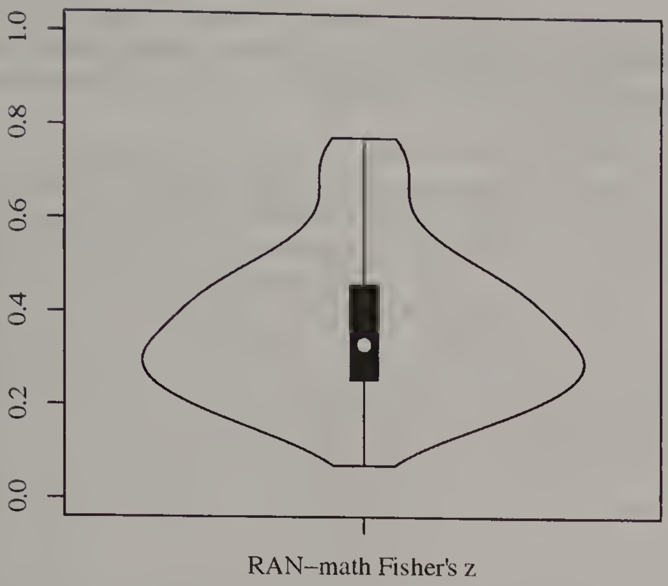


Figure 4. Violin plot of Fisher's z transformed correlation coefficients between RAN and Math (N = 151).

the present study it was not possible to compare correlations in children with difficulties in math and/or naming to those of typically achieving children. Finally, it was expected that RAN would correlate more strongly with mathematics in concurrent studies than in longitudinal studies. Although there was a trend showing that concurrent correlations are higher than longitudinal correlations, the difference was not significant.

Limitations

Some limitations of the present study are worth mentioning. The search procedure may have left out some relevant reports. For example, only published studies, theses, and dissertations were included, which is known to cause some publication bias, meaning that studies with large and/or statistically significant effects, relative to reports with small or null effects, are more likely to be published (Polanin, Tanner-Smith, & Hennessy, 2016). However, publication bias in general may be less important when looking at correlational studies when the effect sizes are purely descriptive and not related to the outcome of the study (e.g., intervention effect). Heterogeneity among the RAN–mathematics correlations was anticipated, but it was larger than expected. In this study, we used independent and correlated correlation coefficients to develop a general understanding regarding the RAN–math association. This means that the independence of observation assumption was purposefully violated in the overall association calculation. However, we used the robust variance estimation method in the moderator analysis that used correlated data. Due to the low number of unique samples, only one partitioned moderator was analyzed at a time and it was not possible to analyze interactions between moderator factors. Some moderator effects might be confounded (e.g., math achievement was mostly assessed with accuracy-based measurements). In addition, much of the variance in the RAN–mathematics correlations was left unexplained. A possible reason for this could be that there are interactions between some of the moderators that were not taken into account. For example, math age could matter only if arithmetic fluency (not math accuracy) was used as an outcome measure. Finally, we used Fisher's z values in the calculations, which may bias the results upward (cf., Schmidt & Hunter, 2015). Also, we followed the metaregression approach for correlated data in which the moderators were dummy coded; however, there is still relatively little research on the method itself and its biases (Hedges et al., 2010; Tipton, 2015). Forthcoming research may better understand the heterogeneity of the RAN–mathematics relationship and moderator associations within a meta-analytic structural equation modeling framework (Cheung, 2008, 2015a, 2015b). More studies are also needed in order to analyze possible interaction effects between the different moderators.

Conclusions and Implications

To date, we have learned about RAN through the studies examining its value as a predictor of reading (e.g., Araújo et al., 2015). The current study revealed that RAN is also a strong correlate of mathematics (particularly of math fluency). A practical implication of this finding is that RAN could be used as an early predictor of mathematics development and perhaps even as a risk factor of future mathematics difficulties, particularly of arithmetic dysfluency (e.g., Koponen, Aro, Räsänen, & Ahonen, 2007; Koponen, Mononen, Räsänen, & Ahonen, 2006; Waber, Wolff, Forbes, & Weiler, 2000). This should be taken into account when assessing school readiness, monitoring skill development, and planning for educational support. In addition, given that no significant differences were found in the size of the correlations between the different types of RAN tasks and mathematics, researchers may rely on nonalphanumeric RAN as a predictor of mathematics performance even before children go to school and become familiar with letters and digits.

From a theoretical point of view, the effects of the moderators (arithmetic vs. math achievement, single-digit vs. multidigit, and fluency vs. accuracy in math) support the view that the RAN–mathematics relationship can at least partially be explained by shared underlying processing requirements, that is, rapid access from visual stimuli to phonological representation stored in long-term memory. Previous studies suggested that this kind of process is important in fluent calculation (de Smedt & Boets, 2010; Koponen et al., 2013). Equally strong correlations between nonalphanumeric RAN, alphanumeric RAN, and mathematics suggest that the relationship between RAN and mathematics is related to both conceptual and phonological processing factors.

References

Ackerman, P. T., Holloway, C. A., Youngdahl, P. L., & Dykman, R. A. (2001). The double-deficit theory of reading disability does not fit all. *Learning Disabilities Research & Practice, 16,* 152–160. http://dx.doi.org/10.1111/0938-8982.00016

Adler, D. (2005). *Vioplot package* [Software]. Retrieved from https://cran.r-project.org/web/packages/vioplot/vioplot.pdf

Araújo, S., Reis, A., Petersson, K. M., & Faísca, L. (2015). Rapid automatized naming and reading performance: A meta-analysis. *Journal of Educational Psychology, 107,* 868–883. http://dx.doi.org/10.1037/edu0000006

Berg, D. H. (2008). Working memory and arithmetic calculation in children: The contributory roles of processing speed, short-term memory, and reading. *Journal of Experimental Child Psychology, 99,* 288–308. http://dx.doi.org/10.1016/j.jecp.2007.12.002

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. http://dx.doi.org/10.1002/9780470743386

Bowey, J. A., McGuigan, M., & Ruschena, A. (2005). On the association between serial naming speed for letters and digits and word-reading skill: Towards a developmental account. *Journal of Research in Reading, 28,* 400–422. http://dx.doi.org/10.1111/j.1467-9817.2005.00278.x

Butterworth, B. (2005). Developmental dyscalculia. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 455–467). Hove, UK: Psychology Press.

Chan, B. M. Y., & Ho, C. S. H. (2010). The cognitive profile of Chinese children with mathematics difficulties. *Journal of Experimental Child Psychology, 107,* 260–279. http://dx.doi.org/10.1016/j.jecp.2010.04.016

Cheung, M. W.-L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. *Psychological Methods, 13,* 182–202. http://dx.doi.org/10.1037/a0013163

Cheung, M. W-L. (2015a). *Meta-analysis. A structural equation modeling approach*. Chichester, West Sussex: Wiley.

Cheung, M. W-L. (2015b). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5,* 1521. http://dx.doi.org/10.3389/fpsyg.2014.01521

Cirino, P. T. (2011). The interrelationships of mathematical precursors in kindergarten. *Journal of Experimental Child Psychology, 108,* 713–733. http://dx.doi.org/10.1016/j.jecp.2010.11.004

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Compton, D. L. (2003). Modeling the relationship between growth in rapid naming speed and growth in decoding skill in first-grade children. *Journal of Educational Psychology, 95,* 225–239. http://dx.doi.org/10.1037/0022-0663.95.2.225

Cox, N. J. (2008). Speaking Stata: Correlations with confidence, or Fisher's z revisited. *The Stata Journal, 8,* 413–439.

Cummine, J., Szepesvari, E., Chouinard, B., & Georgiou, G. (2015). An examination of the rapid automatized naming–reading relationship using functional magnetic resonance imaging. *Neuroscience, 305,* 49–66. http://dx.doi.org/10.1016/j.neuroscience.2015.07.071

Cummine, J., Szepesvari, E., Chouinard, B., Hanif, W., & Georgiou, G. K. (2014). A functional investigation of RAN letters, digits, and objects: How similar are they? *Behavioural Brain Research, 275,* 157–165. http://dx.doi.org/10.1016/j.bbr.2014.08.038

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20,* 487–506. http://dx.doi.org/10.1080/02643290244000239

de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology, 91,* 450–476. http://dx.doi.org/10.1037/0022-0663.91.3.450

de Jong, P. F., & van der Leij, A. (2002). Effects of phonological abilities and linguistic comprehension on the development of reading. *Scientific Studies of Reading, 6,* 51–77. http://dx.doi.org/10.1207/S1532799XSSR0601_03

De Smedt, B., & Boets, B. (2010). Phonological processing and arithmetic fact retrieval: Evidence from developmental dyslexia. *Neuropsychologia, 48,* 3973–3981. http://dx.doi.org/10.1016/j.neuropsychologia.2010.10.018

De Smedt, B., Taylor, J., Archibald, L., & Ansari, D. (2010). How is phonological processing related to individual differences in children's arithmetic skills? *Developmental Science, 13,* 508–520. http://dx.doi.org/10.1111/j.1467-7687.2009.00897.x

Donker, M., Kroesbergen, E. H., Slot, E. M., Van Viersen, S., & De Bree, E. H. (2016). Alphanumeric and non-alphanumeric Rapid automatized naming in children with reading and/or spelling difficulties and mathematical difficulties. *Learning and Individual Differences, 47,* 80–87. http://dx.doi.org/10.1016/j.lindif.2015.12.011

Eklund, K. M., Torppa, M., & Lyytinen, H. (2013). Predicting reading disability: Early cognitive risk and protective factors. *Dyslexia: An International Journal of Research and Practice, 19,* 1–10. http://dx.doi.org/10.1002/dys.1447

Fisher, Z., & Tipton, E. (2015). *Package 'robumeta'* [Software]. Retrieved from https://cran.r-project.org/web/packages/robumeta/robumeta.pdf

Foster, M. E., Jason, L. A., Clements, D. H., & Sarama, J. H. (2015). Processes in the development of mathematics in kindergarten children from Title 1 schools. *Journal of Experimental Child Psychology, 140,* 56–73. http://dx.doi.org/10.1016/j.jecp.2015.07.004

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97,* 493–513. http://dx.doi.org/10.1037/0022-0663.97.3.493

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology, 47,* 1539–1552. http://dx.doi.org/10.1037/a0025510

Georgiou, G., Torppa, M., Manolitsis, G., Lyytinen, H., & Parrila, R. (2012). Longitudinal predictors of reading and spelling across languages varying in orthographic consistency. *Reading and Writing, 25,* 321–346. http://dx.doi.org/10.1007/s11145-010-9271-x

Georgiou, G. K., Tziraki, N., Manolitsis, G., & Fella, A. (2013). Is rapid automatized naming related to reading and mathematics for the same reason(s)? A follow-up study from kindergarten to Grade 1. *Journal of Experimental Child Psychology, 115,* 481–496. http://dx.doi.org/10.1016/j.jecp.2013.01.004

Hannula, M. M., Lepola, J., & Lehtinen, E. (2010). Spontaneous focusing on numerosity as a domain-specific predictor of arithmetical skills. *Journal of Experimental Child Psychology, 107,* 394–406. http://dx.doi.org/10.1016/j.jecp.2010.06.004

Hart, S. A., Petrill, S. A., Thompson, L. A., & Plomin, R. (2009). The ABCs of math: A genetic analysis of mathematics and its links with reading ability and general cognitive ability. *Journal of Educational Psychology, 101,* 388–402. http://dx.doi.org/10.1037/a0015115

Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computation skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology, 79,* 192–227. http://dx.doi.org/10.1006/jecp.2000.2586

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1,* 39–65. http://dx.doi.org/10.1002/jrsm.5

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558. http://dx.doi.org/10.1002/sim.1186

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician, 52,* 181–184.

Humphreys, G. W., Price, C. J., & Riddoch, M. J. (1999). From objects to names: A cognitive neuroscience approach. *Psychological Research, 62,* 118–130. http://dx.doi.org/10.1007/s004260050046

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology, 85,* 103–119.

Juul, H., Poulsen, M., & Elbro, C. (2014). Separating speed from accuracy in beginning reading development. *Journal of Educational Psychology, 106,* 1096–1106. http://dx.doi.org/10.1037/a0037100

Kirby, J. K., Parrila, R. K., & Pfeiffer, S. L. (2003). Naming speed and phonological awareness as predictors of reading development. *Journal of Educational Psychology, 95,* 453–464. http://dx.doi.org/10.1037/0022-0663.95.3.453

Kleemans, T., Segers, E., & Verhoeven, L. (2012). Naming speed as a clinical marker in predicting basic calculation skills in children with specific language impairment. *Research in Developmental Disabilities, 33,* 882–889. http://dx.doi.org/10.1016/j.ridd.2011.12.007

Koponen, T., Aro, T., Räsänen, P., & Ahonen, T. (2007). Language-Based Retrieval Difficulties in Arithmetic: A single case intervention study comparing two children with SLI. *Educational & Child Psychology, 24,* 98–107.

Koponen, T., Aunola, K., Ahonen, T., & Nurmi, J.-E. (2007). Cognitive predictors of single-digit and procedural calculation skills and their covariation with reading skill. *Journal of Experimental Child Psychology, 97,* 220–241. http://dx.doi.org/10.1016/j.jecp.2007.03.001

Koponen, T., Mononen, R., Räsänen, P., & Ahonen, T. (2006). Basic numeracy in children with specific language impairment: Heterogeneity and connections to language. *Journal of Speech, Language, and Hearing Research, 49,* 58–73. http://dx.doi.org/10.1044/1092-4388(2006/005)

Koponen, T., Salmi, P., Eklund, K., & Aro, T. (2013). Counting and RAN: Predictors of arithmetic calculation and reading fluency. *Journal of Educational Psychology, 105,* 162–175. http://dx.doi.org/10.1037/a0029285

Koponen, T., Salmi, P., Torppa, M., Eklund, K., Aro, T., Aro, M., . . . Nurmi, J.-E. (2016). Counting and rapid naming predict the fluency of arithmetic and reading skills. *Contemporary Educational Psychology, 44–45,* 83–94. http://dx.doi.org/10.1016/j.cedpsych.2016.02.004

Korhonen, T. T. (1995). The persistence of rapid naming problems in children with reading disabilities: A nine-year follow-up. *Journal of Learning Disabilities, 28,* 232–239. http://dx.doi.org/10.1177/002221949502800405

Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction, 19,* 513–526. http://dx.doi.org/10.1016/j.learninstruc.2008.10.002

Landerl, K., Fussenegger, B., Moll, K., & Willburger, E. (2009). Dyslexia and dyscalculia: Two learning disorders with different cognitive profiles. *Journal of Experimental Child Psychology, 103,* 309–324. http://dx.doi.org/10.1016/j.jecp.2009.03.006

Landerl, K., & Willburger, E. (2010). Temporal processing, attention, and learning disorders. *Learning and Individual Differences, 20,* 393–401. http://dx.doi.org/10.1016/j.lindif.2010.03.008

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology, 100,* 150–161. http://dx.doi.org/10.1037/0022-0663.100.1.150

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology: General, 124,* 83–97.

Lepola, J., Niemi, P., Kuikka, M., & Hannula, M. M. (2005). Cognitive-linguistic skills and motivation as longitudinal predictors of reading and arithmetic achievement: A follow-up study from kindergarten to Grade 2. *International Journal of Educational Research, 43,* 250–271. http://dx.doi.org/10.1016/j.ijer.2006.06.005

Lervåg, A., Bråten, I., & Hulme, C. (2009). The cognitive and linguistic foundations of early reading development: A Norwegian latent variable longitudinal study. *Developmental Psychology, 45,* 764–781. http://dx.doi.org/10.1037/a0014132

Liao, C. H., Georgiou, G., & Parrila, R. (2008). Rapid naming speed and Chinese character recognition. *Reading and Writing, 21,* 231–253. http://dx.doi.org/10.1007/s11145-007-9071-0

Lipsey, M. W., & Wilson, D. B. (2001). *Applied social research methods: Practical meta-analysis.* Thousand Oaks, CA: Sage.

Manis, F. R., Seidenberg, M. S., & Doi, L. M. (1999). See Dick RAN: Rapid naming and the longitudinal prediction of reading subskills in first and second graders. *Scientific Studies of Reading, 3,* 129–157. http://dx.doi.org/10.1207/s1532799xssr0302_3

Mazzocco, M. M. M., & Grimm, K. J. (2013). Growth in rapid automatized naming from grades K to 8 in children with math or reading disabilities. *Journal of Learning Disabilities, 46,* 517–533. http://dx.doi.org/10.1177/0022219413477475

McBride-Chang, C., & Manis, F. R. (1996). Structural invariance in the associations of naming speed, phonological awareness, and verbal reasoning in good and poor readers: A test of the double deficit hypothesis. *Reading and Writing, 8,* 323–339.

Meyer, M. S., Wood, F. B., Hart, L. A., & Felton, R. H. (1998). Selective prodective value of radpi automatized naming in poor readers. *Jourmal of Learning Disabilities, 31,* 106–117.

Misra, M., Katzir, T., Wolf, M., & Poldrack, R. A. (2004). Neural systems for rapid automatized naming in skilled readers: Unraveling the RAN-reading relationship. *Scientific Studies of Reading, 8,* 241–256. http://dx.doi.org/10.1207/s1532799xssr0803_4

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology, 62,* 1006–1012. http://dx.doi.org/10.1016/j.jclinepi.2009.06.005

Moll, K., Göbel, S. M., & Snowling, M. J. (2015). Basic number processing in children with specific learning disorders: Comorbidity of reading and mathematics disorders. *Child Neuropsychology, 21,* 399–417. http://dx.doi.org/10.1080/09297049.2014.899570

Niklas, F., & Schneider, W. (2013). Home literacy environment and the beginning of reading and spelling. *Contemporary Educational Psychology, 38,* 40–50. http://dx.doi.org/10.1016/j.cedpsych.2012.10.001

Niklas, F., & Schneider, W. (2014). Casting the die before the die is cast: The importance of the home numeracy environment for preschool children. *European Journal of Psychology of Education, 29,* 327–345. http://dx.doi.org/10.1007/s10212-013-0201-6

Norton, E. S., & Wolf, M. (2012). Rapid Automatized Naming (RAN) and Reading Fluency: Implications for Understanding and Treatment of Reading Disabilities. *Annual Review of Psychology, 63,* 427–452. http://dx.doi.org/10.1146/annurev-psych-120710-100431

Parrila, R., Kirby, J. R., & McQuarrie, L. (2004). Articulation rate, naming speed, verbal short-term memory, and phonological awareness: Longitudinal predictors of early reading development? *Scientific Studies of Reading, 8,* 3–26. http://dx.doi.org/10.1207/s1532799xssr0801_2

Polanin, J. R., Tanner-Smith, E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research, 86,* 207–236. http://dx.doi.org/10.3102/0034654315582067

Price, C. J., & McCrory, E. (2005). Functional brain imaging studies of skilled reading and developmental dyslexia. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 473–496). Oxford, UK: Blackwell Publishing. http://dx.doi.org/10.1002/9780470757642.ch25

Rousselle, L., & Noël, M. P. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic vs non-symbolic number magnitude processing. *Cognition, 102,* 361–395. http://dx.doi.org/10.1016/j.cognition.2006.01.005

R Foundation. (2015). *The R project for statistical computing* [Software]. Retrieved from https://www.r-project.org/

Savage, R., & Frederickson, N. (2005). Evidence of a highly specific relationship between rapid automatic naming of digits and text-reading speed. *Brain and Language, 93,* 152–159. http://dx.doi.org/10.1016/j.bandl.2004.09.005

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75–119). Timonium, MD: York Press.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis. Correcting error and bias in research findings.* Thousand Oaks, CA: Sage.

Simmons, F. R., & Singleton, C. (2008). Do weak phonological representations impact on arithmetic development? A review of research into

arithmetic and dyslexia. *Dyslexia: An International Journal of Research and Practice, 14,* 77–94. http://dx.doi.org/10.1002/dys.341

Song, S., Georgiou, G., Su, M.-M., & Shu, H. (2016). How well do phonological awareness and rapid automatized naming correlate with reading accuracy and fluency in Chinese? *A meta-analysis. Scientific Studies of Reading, 20,* 99–123. http://dx.doi.org/10.1080/10888438.2015.1088543

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 99–110). Chichester, UK: Wiley. http://dx.doi.org/10.1002/0470870168.ch6

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., . . . Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 343,* d4002. http://dx.doi.org/10.1136/bmj.d4002

Swanson, H. L. (2006a). Cognitive processes that underlie mathematical precociousness in young children. *Journal of Experimental Child Psychology, 93,* 239–264. http://dx.doi.org/10.1016/j.jecp.2005.09.006

Swanson, H. (2006b). Cross-sectional and incremental changes in working memory and mathematical problem solving. *Journal of Educational Psychology, 98,* 265–281. http://dx.doi.org/10.1037/0022-0663.98.2.265

Swanson, H. L. (2008). Working memory and intelligence in children: What develops? *Journal of Educational Psychology, 100,* 581–602. http://dx.doi.org/10.1037/0022-0663.100.3.581

Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 96,* 471–491. http://dx.doi.org/10.1037/0022-0663.96.3.471

Swanson, H. L., Jerman, O., & Zheng, X. (2008). Growth in working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 100,* 343–379. http://dx.doi.org/10.1037/0022-0663.100.2.343

Swanson, H. L., & Kim, K. (2007). Working memory, short-term memory, and naming speed as predictors of children's mathematical performance. *Intelligence, 35,* 151–168. http://dx.doi.org/10.1016/j.intell.2006.07.001

Swanson, H. I., Trainin, G., Necoechea, D. M., & Hammill, D. D. (2003). Rapid naming, phonological awareness, and reading: A meta-analysis of the correlation evidence. *Review of Educational Research, 73,* 407–440. http://dx.doi.org/10.3102/00346543073004407

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5,* 13–30. http://dx.doi.org/10.1002/jrsm.1091

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20,* 375–393. http://dx.doi.org/10.1037/met0000011

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Burgess, S., & Hecht, S. (1997). Contributions of phonological awareness and automatic naming ability to the growth of word-reading skills in second- to fifth-grade children. *Scientific Studies of Reading, 1,* 161–185. http://dx.doi.org/10.1207/s1532799xssr0102_4

Träff, U. (2013). The contribution of general cognitive abilities and number abilities to different aspects of mathematics in children. *Journal of Experimental Child Psychology, 116,* 139–156. http://dx.doi.org/10.1016/j.jecp.2013.04.007

van Bergen, E., de Jong, P. F., Maassen, B., & van der Leij, A. (2014). The effect of parents' literacy skills and children's preliteracy skills on the risk of dyslexia. *Journal of Abnormal Child Psychology, 42,* 1187–1200. http://dx.doi.org/10.1007/s10802-014-9858-9

van Daal, V., van der Leij, A., & Adèr, H. (2013). Specificity and overlap in skills underpinning reading and arithmetical fluency. *Reading and Writing, 26,* 1009–1030. http://dx.doi.org/10.1007/s11145-012-9404-5

van der Sluis, S., de Jong, P. F., & van der Leij, A. (2007). Executive functioning in children, and its relations with reasoning, reading, and arithmetic. *Intelligence, 35,* 427–449. http://dx.doi.org/10.1016/j.intell.2006.09.001

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metaphor package. *Journal of Statistical Software, 36,* 1–48. http://dx.doi.org/10.18637/jss.v036.i03

Waber, D. P., Wolff, P. H., Forbes, P. W., & Weiler, M. D. (2000). Rapid automatized naming in children referred for evaluation of heterogeneous learning problems: How specific are naming speed deficits to reading disability? *Child Neuropsychology, 6,* 251–261. http://dx.doi.org/10.1076/chin.6.4.251.3137

Wei, T. Q., Bi, H. Y., Chen, B. G., Liu, Y., Weng, X. C., & Wydell, T. N. (2014). Developmental changes in the role of different metalinguistic awareness skills in Chinese reading acquisition from preschool to third grade. *PLoS ONE, 9*(5), e96240. http://dx.doi.org/10.1371/journal.pone.0096240

Willburger, E., Fussenegger, B., Moll, K., Wood, G., & Landerl, K. (2008). Naming speed in dyslexia and dyscalculia. *Learning and Individual Differences, 18,* 224–236. http://dx.doi.org/10.1016/j.lindif.2008.01.003

Wimmer, H., Mayringer, H., & Landerl, K. (1998). Poor reading: A deficit in skill-automatization or a phonological deficit? *Scientific Studies of Reading, 2,* 321–340. http://dx.doi.org/10.1207/s1532799xssr0204_2

Wocadlo, C., & Rieger, I. (2007). Phonology, rapid naming and academic achievement in very preterm children at eight years of age. *Early Human Development, 83,* 367–377. http://dx.doi.org/10.1016/j.earlhumdev.2006.08.001

Appendix

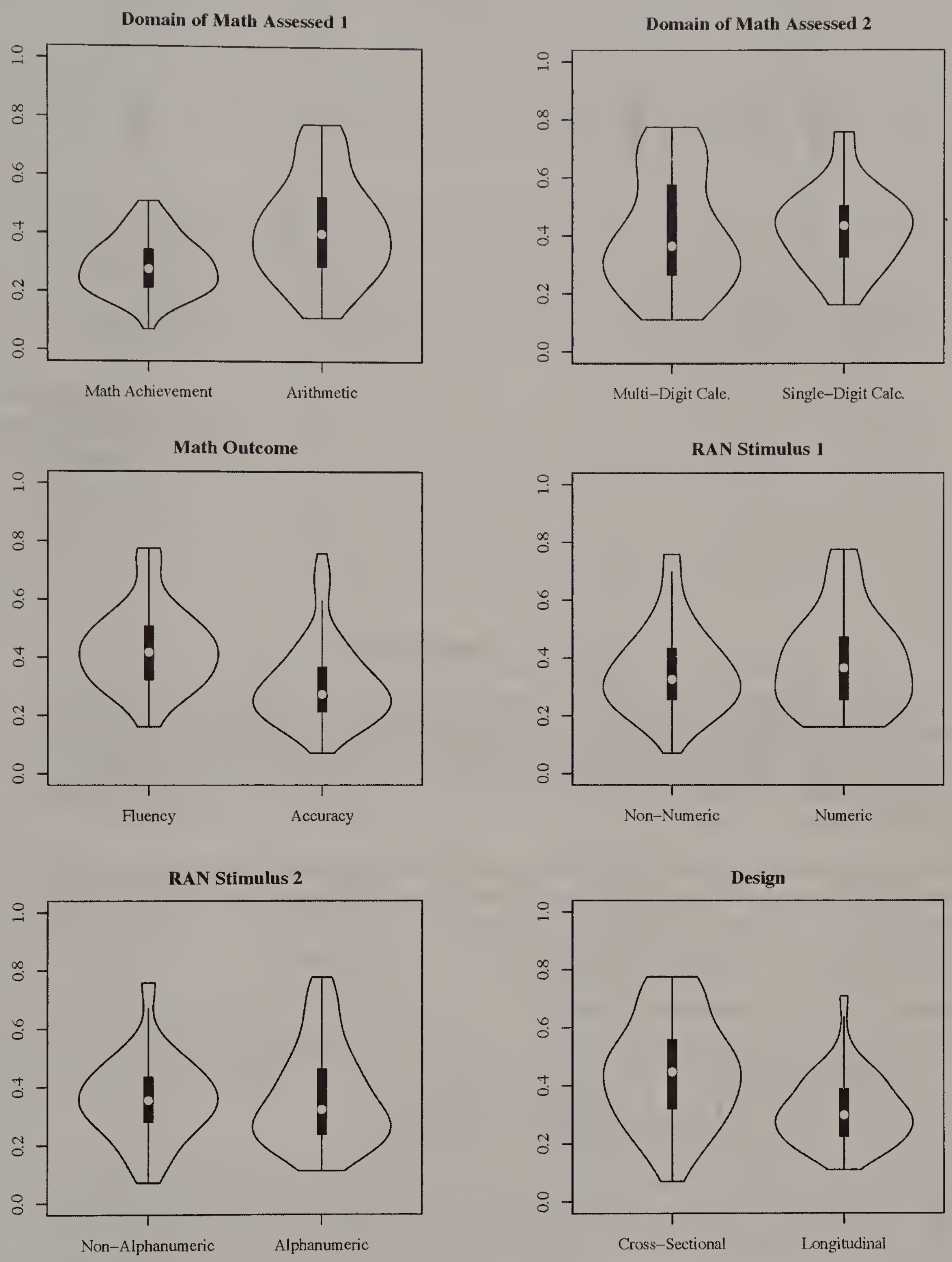


Figure A1. Violin plots of Fisher's z transformed correlation coefficients for procedural moderators.

(*Appendix continues*)

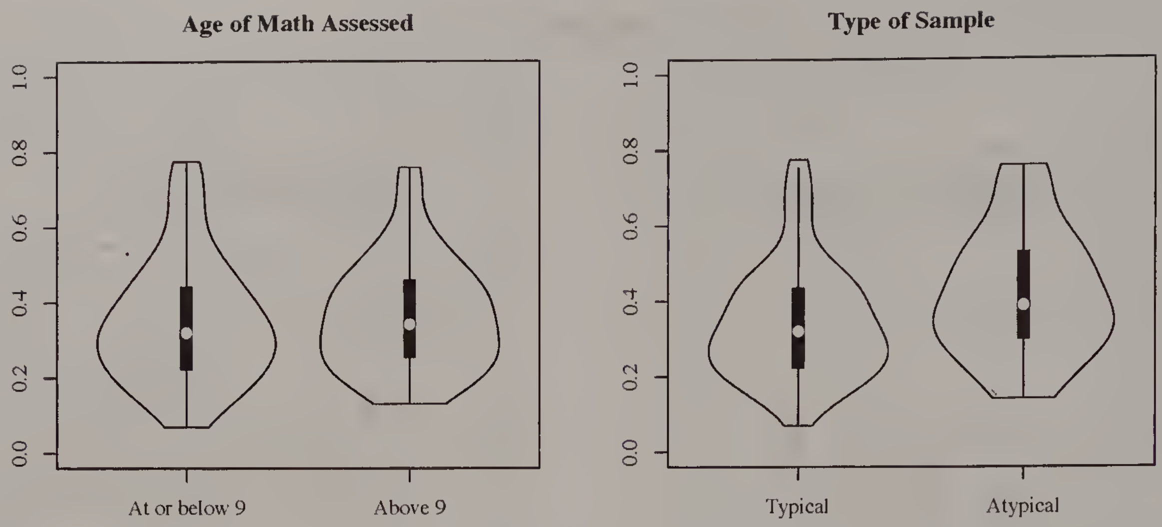


Figure A2. Violin plots of Fisher's z transformed correlation coefficients for sample moderators.

Received April 2, 2016
Revision received November 16, 2016
Accepted December 2, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 7, 993–1009

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000183

Maximizing Gender Equality by Minimizing Course Choice Options? Effects of Obligatory Coursework in Math on Gender Differences in STEM

Nicolas Hübner, Eike Wille, Jenna Cambria, Kerstin Oschatz, Benjamin Nagengast, and Ulrich Trautwein
Hector Research Institute of Education Sciences and Psychology, University of Tübingen

Math achievement, math self-concept, and vocational interests are critical predictors of STEM careers and are closely linked to high school coursework. Young women are less likely to choose advanced math courses in high school, and encouraging young women to enroll in advanced math courses may therefore bring more women into STEM careers. We looked at a German statewide educational reform that required all students to take advanced math courses and examined differential effects of the reform on young men and women's math achievement, math self-concept, vocational interests, and field of study at university. We compared data from 4,730 students before the reform and 4,715 students after the reform. We specified multiple regression models and tested main effects of gender and cohort as well as the effect of the Cohort × Gender interaction on all outcomes. All outcomes showed clear gender differences favoring young men before the reform. However, the reform was associated with different effects for young men and women: Whereas gender differences in math achievement were smaller after the reform, differences between young men and women in math self-concept and realistic and investigative vocational interests were larger after the reform than before. Gender differences in the field of study at university did not differ between before and after the reform. Results suggest that reducing course choice options in high school does not automatically increase gender equality in STEM fields.

Educational Impact and Implications Statement

This study suggests that making it obligatory for young women and men to participate in advanced math courses in upper secondary school can increase their math achievement and realistic (e.g., technical) interests. However, it also seems to have the potential to negatively impact young women's self-perceptions of their math ability. The study illustrates that well-intended educational reforms might not achieve all goals (and in fact might result in unintended side effects) when psychological factors are ignored.

Keywords: gender differences, school reform, math achievement, math self-concept, vocational interests

Women are underrepresented in mathematically intensive STEM (science, technology, engineering, and mathematics) domains (Ceci, Williams, & Barnett, 2009; Schoon & Eccles, 2014). Gender disparities in STEM fields are crucial for the larger economy because the presence of more women would diversify the workforce and might add to a more competitive work environment with an increased number of qualified employees in this area (e.g., National Science Foundation [NSF], 2013; OECD, 2010). In addition, women's underrepresentation also matters to gender inequity in income because STEM fields provide high-status career options (e.g., Sells, 1980; Watt, Eccles, & Durik, 2006). Advanced high school coursework in math is a key predictor of STEM career choices (Ma & Johnson, 2008), and young women are less likely to choose advanced math courses than young men (Nagy et al., 2008; Updegraff, Eccles, Barber, & O'Brien, 1996). Thus, it is important to ask whether the challenge of recruiting more women into STEM careers may be addressed by mandatory enrollment in advanced math courses in high school (e.g., by changing course assignment procedures; Ma & Johnson, 2008; Sells, 1980). However, there is limited real-world data on the effectiveness of such reforms.

This article was published Online First March 27, 2017.

Nicolas Hübner, Eike Wille, Jenna Cambria, Kerstin Oschatz, Benjamin Nagengast, and Ulrich Trautwein, Hector Research Institute of Education Sciences and Psychology, University of Tübingen.

Jenna Cambria is now at the University of Arkansas.

The first two authors contributed equally to this work and are listed in alphabetical order. This research was partly funded by the state government of Baden-Württemberg. Nicolas Hübner and Eike Wille are doctoral students at the LEAD Graduate School & Research Network (GSC 1028), funded by the Excellence Initiative of the German federal and state governments. We thank Steve West for helpful feedback on a previous version of this article.

Correspondence concerning this article should be addressed to Nicolas Hübner and Eike Wille, Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Europastraße 6, 72072 Tübingen, Germany. E-mail: nicolas.huebner@uni-tuebingen.de or eike.wille@uni-tuebingen.de

In the present study, we reanalyzed representative data from a large school achievement study on the effects of a major reform of upper secondary education in a large state in Germany. More specifically, the reform required all students to take an advanced math course, which successfully eliminated a prior imbalance between young men and women in these advanced courses. We studied the effects of this school reform on gender differences in math achievement, math self-concept, and interests in realistic and investigative areas because such outcomes are critical in terms of later educational choices. Furthermore, we investigated effects on students' actual field of study at university 2 years after they completed high school.

Predictors of Gendered Career Choices in STEM

Academic Achievement and STEM Career Choices

In explaining STEM career choices for young men and women, research on educational choices has traditionally focused on the role of math achievement on career interests (e.g., Parker et al., 2012; Sells, 1980). Such work has consistently shown that math achievement is a key predictor of both high school subject choices and later career choices, particularly with respect to mathematically intensive STEM careers (Parker et al., 2012; Sells, 1980). For instance, there is evidence that high school math achievement predicts career aspirations in STEM during high school (e.g., Ma & Johnson, 2008), field of study at university (e.g., Parker et al., 2012), and university retention (Alarcon & Edwards, 2013).

The relation between academic achievement and career choice is often explained by employing rational choice models (Gottfredson, 1986; Lubinski & Benbow, 2006). First, individuals prefer careers that provide activities they expect to be good at. Second, individuals who have the required competencies gain access to the professional field, for instance, due to admission restrictions for college majors. Third, individuals tend to leave professions if their competencies are insufficient for the specific profession. Thus, young people with high math achievement have a tendency to pursue mathematically intensive STEM careers such as physics, engineering, or informatics (Humphreys & Yao, 2002; Parker et al., 2012).

Self-Concept and STEM Career Choices

Above and beyond the effects of achievement, young people's career choices are also critically linked to their academic self-concept in high school (Schoon & Eccles, 2014; Watt & Eccles, 2008). Academic self-concept is defined as a person's self-evaluation of his or her own general ability in a specific domain, such as doing well in math (Bong & Skaalvik, 2003; Marsh, 1986). In developing a domain-specific self-concept, students refer to their own achievement in a domain but also compare their own ability with their interpretation of peers' achievements in the same domain (e.g., Marsh, 1986; Marsh et al., 2015).

In fact, self-concept has been shown to be related to future-oriented motivation and aspirations such as career choices (e.g., Schoon & Eccles, 2014; Watt & Eccles, 2008); math self-concept has been identified as positively related to various educational outcomes in the STEM area, such as high school students' educational aspirations within the STEM fields (Jansen, Scherer, & Schroeders, 2015; Schoon & Eccles, 2014) and choice and retention of mathematically intensive STEM university subjects (Perez, Cromley, & Kaplan, 2014; Schoon & Eccles, 2014) for both men and women.

It is important to mention that self-concept does not measure the same thing as self-efficacy, although they are closely related (e.g., Bong & Skaalvik, 2003). Furthermore, self-concept predicts educational biographies and trajectories, whereas self-efficacy is used for predicting success in a specific task (Jansen et al., 2015).

Vocational Interests and STEM Career Choices

Next to math achievement and self-concept, vocational interests are very important in predicting STEM career choices. The role of interest for achievement-related outcomes is well-established (Schoon & Eccles, 2014; Su, Rounds, & Armstrong, 2009). Whereas educational psychology has traditionally focused on children's and adolescents' interest in learning and achievement in the school context (Krapp, 1999; Wigfield & Cambria, 2010), research and theories in vocational psychology, such as Holland's theory of vocational interests (Holland, 1959, 1997), have been highly effective at addressing young people's postschool career choices with interests describing activities in fields of professions or university majors (Rounds & Su, 2014; Su & Rounds, 2015). Vocational interests are central predictors of vocational choices such as the selection of a college major or profession (Humphreys & Yao, 2002; Pässler, Beinicke, & Hell, 2014) and are also crucial for job performance and turnover (Nye, Su, Rounds, & Drasgow, 2012) as well as income (Huang & Pearce, 2013).

Holland (1966) defined vocational interests as "the expression of personality in work, hobbies, recreational activities, and preferences" (p. 3) and expected that they would directly influence goal-oriented behaviors. He posited that individuals should strive for educational and occupational environments that are in line with their interests, and there is a large body of research that supports this proposition (e.g., Humphreys & Yao, 2002; Strong, 1943). Vocational interests are therefore defined as trait-like preferences for activities, and these preferences are captured on a very general level (Holland, 1997; Rounds & Su, 2014). In this regard, vocational interests differ from the term *interest* in educational psychology. Interest in educational psychology is usually defined as a motivational variable that "refers to the psychological state of engaging or the predisposition to reengage" (Hidi & Renninger, 2006, p. 112). Contrary to conceptualizations of interest in educational psychology, which usually focus on domain-specific interest in single (school) subjects (e.g., Hidi & Ainley, 2002), vocational interests emphasize broad sets of activities and experiences that go with different kinds of professions. Thereby, Holland's model represents six interest domains, which describe activities that are related to different careers: *realistic*, *investigative*, *artistic*, *social*, *enterprising*, and *conventional*. In our study, we focused on the realistic and investigative dimensions because they have been shown to be related to mathematically intensive STEM fields (Ackerman & Heggestad, 1997; Su et al., 2009). People with high realistic interests tend to like working with things and prefer activities that involve the manipulation of objects, tools, and machines. People with high investigative interests are likely to be interested in understanding how physical and biological phenomena function and tend to prefer activities that include analyzing and

problem solving on a more abstract level (Holland, 1997). Consequently, young people with realistic and investigative interests are likely to choose mathematically intensive STEM careers such as physics, engineering, or informatics (Su & Rounds, 2015; Su et al., 2009).

Gender Differences in Math Achievement, Math Self-Concept, and Realistic and Investigative Interests

Gender differences in math achievement have often been used to explain gendered career choices in the STEM domains (e.g., Hyde, Fennema, Ryan, Frost, & Hopp, 1990; Reilly, Neumann, & Andrews, 2015). Historically, there has been a pattern of young men outperforming young women in math achievement (e.g., Hyde, Fennema, & Lamon, 1990). However, more recent research has provided mixed evidence: Some studies have suggested no or only slight differences in math achievement between young women and men in high school (e.g., Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Voyer & Voyer, 2014), whereas others have indicated that such differences still exist and that the magnitude of the differences between young men and women varies between countries and according to the educational requirements of the system (e.g., Else-Quest, Hyde, & Linn, 2010; Reilly et al., 2015). For German samples, previous research has consistently indicated that young men still perform better in math in high school than young women (e.g., Else-Quest et al., 2010; Nagy et al., 2008).

Regarding math self-concept, previous research has shown that—after achievement is controlled for—boys tend to report higher math self-concept than girls even in primary school, and such gender differences remain constant across high school (e.g., Marsh & Yeung, 1998; Nagy et al., 2008).

With respect to realistic and investigative interests, previous research has consistently shown that men score higher on both interest dimensions than women (e.g., Lippa, 1998; Su et al., 2009).

Relations Between Achievement, Self-Concept, and Vocational Interests

Academic achievement, the self-evaluation of academic achievement (i.e., self-concept), and interests have been found to be interrelated, which means that, in general, people are interested in and feel competent in domains they are good at. The relations between these constructs have been described in different theoretical frameworks, such as Eccles et al.'s (1983) expectancy-value theory and Lent, Brown, and Hackett's (1994) social cognitive career theory. According to these theories, prior achievement influences an individual's evaluation of his or her achievement (e.g., self-concept), as well as his or her interests in the same domain. A person's interests are furthermore influenced by his or her perception of competence, and both self-concept and interests are believed to predict later achievement. The rationale behind these relations is that individuals who have positive previous achievement-related experiences in one domain will feel more competent and will develop interests in the same domain. Furthermore, if they feel competent and are interested, they will engage more frequently and intensely in tasks and activities related to that domain, and thereby, they will show high levels of persistence and effort. In the end, this leads to better performance in the same domain (Wigfield, Tonks, & Klauda, 2009).

There is a lot of empirical support for such relations between achievement, self-concept, and interests. With respect to the relation between achievement and self-concept, several studies have indicated that achievement and self-concept are positively correlated (e.g., Chen, Yeh, Hwang, & Lin, 2013), and bidirectional relations have been found, indicating that students' prior achievement influences their self-concept and that their self-concept influence their later achievement (for a review, see Marsh, 2007). Furthermore, there is evidence that self-concept predicts changes in interests (Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Wigfield et al., 1997) and that interests and achievement are also interrelated. Thereby, correlation-based research has shown positive relations between achievement and interests for various conceptualizations of interest, such as individual interest (see Schiefele, Krapp, & Winteler, 1992) or task values (see, e.g., Updegraff et al., 1996), but also for vocational interests, where positive correlations between math achievement and realistic as well as investigative interests have been found (Ackerman & Heggestad, 1997; Rolfhus & Ackerman, 1996). Furthermore, self-concept has been found to predict later interests (e.g., Denissen, Zarrett, & Eccles, 2007; Marsh et al., 2005), and a reciprocal relation has been found between interests and achievement (e.g., Denissen et al., 2007; Jansen, Lüdtke, & Schroeders, 2016). However, prior studies have so far focused on subject-specific conceptualizations of interest, and less is known about directional relations between these constructs and realistic and investigative interests.

Effects of Course Level on Achievement, Self-Concept, and Vocational Interests

Students' achievement, self-concept, and vocational interests have been linked to their enrollment in advanced and basic courses in high school (e.g., Köller, Baumert, & Schnabel, 2001; Marsh, 2005). The effects of high school coursework on achievement, self-concept, and interests have been explained by variability in the benefits for and constraints on students taking basic and advanced courses (e.g., Köller et al., 2001; Marsh, 2005). In Germany, as in most school systems in developed countries, students in upper secondary school self-select into basic and advanced math courses, which differ in terms of curricular content and level as well as in class composition (Schnabel, Alfeld, Eccles, Köller, & Baumert, 2002). These differences between advanced and basic coursework have been found to lead to differential effects on students' achievement, self-concept, and interests, after students' previous performance was controlled for (e.g., Köller et al., 2001; Trautwein, Köller, Lüdtke, & Baumert, 2005). Regarding students' academic achievement, course level and achievement have been found to be positively associated; students in advanced courses have typically shown higher achievement at the end of high school than those in basic courses, even after students' prior achievement was taken into account (e.g., Gamoran & Mare, 1989; Köller et al., 2001).

Effects of course level on self-concept and vocational interests are less clear. Regarding self-concept, positive associations have been found between a student's own achievement and his or her self-concept in the same domain, as described in the previous section (Marsh, 1986; Marsh et al., 2015). Thus, students showed higher self-concept in advanced courses than in basic courses in general (Chmielewski, Dumont, & Trautwein, 2013). However,

students tend to compare their own achievement with the perceived achievement of their classmates and consequently judge their own achievement as relatively lower when they are surrounded by students with higher achievement. Therefore, students in advanced courses have shown a lower self-concept than students with comparable achievement in basic courses (Chmielewski et al., 2013; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006).

With respect to vocational interests, research has shown that students in advanced and basic courses differ in their vocational interests because their course choices are based on their vocational interests (Nagy & Husemann, 2010; Patrick, Care, & Ainley, 2011). However, it is less clear if or how course level might also predict vocational interests. First, a positive association has been identified between achievement and vocational interests as described above, on which basis one might speculate that course level in math might positively influence realistic and investigative interests (Ackerman & Heggestad, 1997; Anthoney & Armstrong, 2010). Second, initial findings have indicated effects of the average level of class achievement on students' vocational interests. Cambria, Brandt, Nagengast, and Trautwein (2016) investigated 10th graders' achievement in several domains and their vocational interests. They found that achievement in math was positively associated with realistic and investigative interests, and that students with the same individual math achievement level had higher realistic and investigative interests when they were in a class with a higher mean level of achievement.

To sum up, math achievement, math self-concept, and vocational interests are central predictors of mathematically intensive STEM career choices, and these predictors explain gendered career choices in these fields. The findings regarding gender differences in math achievement have been inconsistent, but a considerable amount of research has shown that young men demonstrate higher math self-concept and STEM-related vocational interests than young women. Furthermore, the existing literature indicates that students' achievement and self-concept in math as well as their STEM-related interests are closely related to high school coursework.

The Present Study

In the present study, we examined the effects of a reform in upper secondary high school on gender differences in central predictors of STEM career choices and students' choice of STEM university subjects by reanalyzing representative data from 9,545 German students. Math high school coursework has been found to be closely linked to achievement, self-concept, and interests in the STEM fields (Nagy et al., 2008; Updegraff et al., 1996), all of which are central predictors of STEM career choices (Ma & Johnson, 2008; Nagy et al., 2008). A lower percentage of young women than men had chosen advanced math courses before the reform took place, but this difference was completely eliminated by the reform because the reform required all students to take advanced math courses. Thus, we expected effects of the reform on gender differences in STEM-related outcomes.

There is ample evidence of such effects of high school coursework on achievement, self-concept, and interests, but previous research has not addressed how gender differences in math achievement, self-concept, and interests as key predictors of STEM career choices may be influenced by requiring all students to enroll in advanced courses in math. The present study takes a major step toward filling this gap by investigating such an educational policy and its effects on women's participation in the STEM fields. We examined how changes in high school coursework are related to gender differences in predictors of STEM career choices and students' subjects of study at university after school. To do so, we evaluated effects of a school reform that was introduced in 2002 in one of the largest German states. The reform included the abolition of different math courses. Before the reform, students had been allowed to take math as either an advanced or a basic course. After the reform, all students had to take an obligatory advanced-level math course (Ministry of Education, Cultural Affairs, Youth, & Sport, Baden-Württemberg, 2002).

Because high school course level tends to predict students' achievement and self-concept, and because young women were less likely than young men to choose advanced courses in math before the reform, we expected that the effects of the reform on these outcomes would differ between the young women and men in the current study. As positive effects of course level on students' achievement have been documented, we hypothesized that gender differences in math achievement would be smaller after the reform (when all young men and women took advanced math courses) compared with before the reform (when more young men than young women had taken advanced math courses). Here, we assume that the smaller gender differences in achievement expected after the reform would be based on the higher achievement of young women after the reform compared with before. Regarding gender differences in math self-concept, we hypothesized that gender differences would be larger after the reform than before. This proposition was based on the finding that course level tends to have negative effects on a student's self-concept, and there was a higher percentage of young men than young women in advanced courses before the reform, whereas all students took advanced courses after the reform. We therefore expected that young women's self-concept would be lower after the reform than before on average, which would lead to greater gender differences in math self-concept. So far, there is less work on effects of high school coursework on vocational interests, and it is therefore not clear whether and how the reform might be related to gender differences in realistic and investigative interests. However, if we were to find similar effects of course level on STEM-related vocational interests as on self-concept and subject-specific interest, we would tentatively expect larger gender differences in realistic and investigative interests after the reform than before.

Because we expected differential reform effects on central predictors of STEM career choices (math achievement, math self-concept, realistic and investigative interests), we did not specify what the effects on the actual choice of STEM university subjects would be.

Method

The Reform of Upper Secondary School in the German School System

Before the reform of upper secondary school education, students in most German states self-selected their courses and were given the choice between math as an advanced course (about 5 hours per week) or a basic course (about 3 hours per week). In total, each student was required to select two advanced courses and typically

six basic courses in different subjects. The individual combination of advanced and basic courses represented an individual profile for each student for all of their upper secondary school trajectories, and students were not able to choose different courses each semester. Beginning in 2002, most German states enacted reforms of their higher secondary education systems and implemented a course program. This program can be characterized by a reduction in the number of options in favor of a higher subject-related average amount of time allocated across all students to specific compulsory core subjects (German, mathematics, and foreign language). In most states, students were no longer able to self-select into different courses from that point in time on but were instead required to take a total of five courses from specific fields (e.g., math, foreign language, science) for a similar amount of time (4 hr per week). Besides these compulsory courses, students had to participate in other courses for a reduced number of hours (2 hr per week; e.g., arts, science, or social studies; Köller, Watermann, Trautwein, & Lüdtke, 2004; Trautwein, Neumann, Nagy, Lüdtke, & Maaz, 2010). To sum up, the two major changes of the reform were (a) an increase in the number of courses that had to be chosen for final examinations in upper secondary school on an advanced course level and (b) written exams in the first four of these courses (instead of the first three).

Description of Study and Sample

Data were drawn from the study Transformation of the Secondary School System and Academic Careers (TOSCA; Köller et al., 2004; Trautwein et al., 2010). The TOSCA study was designed to assess a representative sample of students in the last 4 months of their final year of upper secondary school in one German state (Baden-Württemberg). The data from the first waves of TOSCA 2002 and TOSCA 2006 are representative for all students in the final year of upper secondary school in the state of Baden-Württemberg. We considered data from N = 149 schools in the first wave of the first cohort (TOSCA 2002; N = 4,730; 54.5% female) as well as data from N = 146 schools in the first wave of the second cohort (TOSCA 2006; N = 4,715; 54.1% female). Over the course of the reform, another school type (biotechnological Gymnasium) was introduced. Robustness checks revealed no differences in results when students from this type of school were included versus not included. In our sample, roughly 60% of the students were enrolled in a general higher secondary school, and 40% were in a vocational upper secondary school. The time between the start of the course and our measurement was approximately 1.5 years. The measurement took place right at the end of the course. Data collection was executed by trained research assistants who visited every class and lasted for approximately 1 day per school. The first cohort contains data from students who chose basic and advanced courses in upper secondary high school, whereas the second cohort consists of data from students who all took the obligatory advanced math courses. The data from the two cohorts were drawn from the same schools. In both cohorts, a second assessment took place 2 years after the first measurement point via questionnaires that were sent to the participants. Overall, 80% of all students agreed to participate in the first wave of TOSCA 2002, and 82% of all students agreed to participate in the first wave of TOSCA 2006. At the second assessment, which followed 2 years after the first assessments for TOSCA 2002 and TOSCA 2006, respectively, information was obtained about students' field of study at university from N = 1,741 students from TOSCA 2002 and N = 2,157 from TOSCA 2006 (see Figure 1).

Instruments

Math achievement. The Advanced Mathematics test was based on items from the Third International Mathematics and Science Study (TIMSS; Mullis et al., 1998). According to Mullis et al. (1998), the advanced mathematics test takes into account "current thinking and priorities in the field of mathematics" (p. 284). The advanced mathematics test contained a total of 68 items from the areas of (a) Numbers, Equations, and Functions, (b) Analysis, (c) Geometry, (d) Propositional Logic and Proofs, as well as (e) Probability and Statistics. Most of the items were related to the first area and directly tested competencies from upper secondary school. Approximately two thirds of all of the items were multiple-choice questions, whereas the other items were administered in an open-ended format. A multimatrix design was used to administer the items; therefore, the students did not work on all 68 items but on a subset of items in one of four booklets that contained six different item clusters that were rotated systematically. In order to be able to compare the two different cohorts (TOSCA 2002 and TOSCA 2006), items were scaled by applying item response theory (IRT; Rasch model) to account for the multimatrix design and to test for differential item functioning. As reported by Nagy, Neumann, Trautwein, and Lüdtke (2010), we used five completed data sets with *plausible values* (PVs), which were estimated in Mplus 5.2. These PVs were based on multiply imputed data, which was imputed previously with NORM (Schafer, 1997). As reported by Nagy, Neumann, et al. (2010), the psychometric properties of the test are good (PV reliability TOSCA 2002: .88; PV reliability TOSCA 2006: .90).

Mathematics self-concept. Mathematics self-concept was measured with four items from the Self-Description Questionnaire III (SDQ III; Marsh & O'Neill, 1984; Marsh & Shavelson, 1985; Marsh, 1992), using the German translation by Schwanzer, Trautwein, Lüdtke, and Sydow (2005). The translated items focused on the evaluation of cognitive aspects (e.g., "I was always good in mathematics," e.g., Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007). The scale showed high internal consistency in both samples, TOSCA 2002 (Cronbach's alpha = .89) and TOSCA 2006 (Cronbach's alpha = .90).

Vocational interests. Vocational interests were assessed with the Revised General Interest Structure Test (AIST-R; Bergmann & Eder, 2005), which is based on Holland's (1997) RIASEC model. This instrument categorizes students with regard to six different

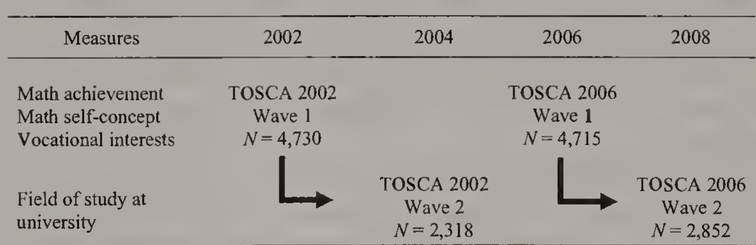


Figure 1. Schematic illustration of the study's timeline. All data in Wave 1 were collected at the end of upper secondary school. TOSCA = Transformation of the Secondary School System and Academic Careers.

dimensions of interest, namely, realistic (R), investigative (I), artistic (A), social (S), enterprising (E), and conventional (C) interests by using a total of 60 items (six 10-item scales). Students were asked to rate how interested they were in the described activities on a 5-point Likert scale ranging from 1 (*not at all*) to 5 (*very much*). An example item of realistic interests is "Working with machines or technical devices" and "Doing physically challenging work," whereas investigative interests were assessed with items such as "Dealing with unexplored things" and "Working in an experimental laboratory." The realistic and investigative facets, which were of specific interest in the present context, showed high internal consistencies (realistic interests—TOSCA 2002: Cronbach's alpha = .86; TOSCA 2006: Cronbach's alpha = .87; investigative interests—TOSCA 2002 and 2006: Cronbach's αs = .85).

Field of study at university. The field of study at university was assessed for each cohort 2 years after they graduated from high school. Students were able to report their subject of study or a combination of study subjects. Students' data were coded according to the official classification system of the Federal Statistical Office, the Fachserie 11 (Federal Statistical Office, 2008). In the current study, we used information about the field of study and computed one variable for which mathematics, engineering, computer science, and physics were coded as STEM subjects only if they were indicated as the first subject of study. In addition, we also specified various alternative codings where only the first, the first two, or all three subject indications were used to calculate the dependent variable and included biology, chemistry, or both as STEM subjects. The general pattern of results was identical across all these different analyses. Furthermore, we did not find any significant differences in STEM-related course change or student withdrawal patterns when comparing the first and second assessments between TOSCA 2002 and TOSCA 2006. The results were based on analyses in which mathematics, engineering, computer science, and physics were coded as STEM subjects.

Covariates. We controlled for the influence of several variables described below.

School types. Because students from different school types (e.g., vocational higher secondary schools and general higher secondary schools) usually differ in cognitive and noncognitive aspects (Trautwein et al., 2010), we included a dummy variable to be able to distinguish between vocational and general higher secondary schools.[1]

Socioeconomic background. Socioeconomic background was measured with information about the highest level of occupation in the family (of either the father or mother) and coded in accordance with the International Standard Classification of Occupations (ISCO-88). The ISCO scores were in turn converted into International Socio economic Index of Occupational Status (ISEI) 88 scores (Ganzeboom, De Graaf, & Treiman, 1992; Ganzeboom & Treiman, 1996). The highest ISEI value between the two parents was used to characterize the socioeconomic background of the students.

Number of books available in the home. The number of books available in the home was measured on a 7-point scale ranging from *zero books available* to *more than 500 books available*. This variable has been shown to be a good indicator of a family's cultural capital (e.g., Evans, Kelley, Sikora, & Treiman, 2010).

Age. The age of the students at the time of the assessment was calculated on the basis of information about students' year and month of birth.

Immigration background. Students with at least one parent born outside of Germany were coded as students with an immigration background.

Statistical Analyses

In order to test for reform effects, we specified multiple regression models involving the TOSCA study survey weights and tested gender as a moderator of the effect of the reform on the different STEM-related outcomes. The models contained the variables gender and cohort as well as socioeconomic background (HISEI), cultural capital (number of books), immigration background, type of school (general Gymnasium vs. type of vocational Gymnasium), and age as covariates. We controlled for these covariates to eliminate the influence of these potential confounders and to increase the precision of our estimation. In addition, we added the Cohort × Gender interaction in order to examine whether the reform had differential effects on young women and men. Because students from different types of schools usually differ in their cognitive and behavioral outcomes (Trautwein et al., 2010), we also controlled for this differential impact by including the three-way interaction between Cohort × School Type × Gender as well as the interaction between School Type × Cohort.

We also specified a multivariate model with a Wald test for the interaction effects and controlled for the false discovery rate of all parameter estimates in each multiple regression afterward by applying the Benjamini-Hochberg adjustment (Benjamini & Hochberg, 1995).

We additionally investigated students' actual field of study at university 2 years after they completed high school. Of special interest in the current analysis were potential differences with regard to whether or not students chose a STEM-related field of study. We therefore specified models to predict field of study in STEM versus other fields of study in multiple logistic regressions.

We used the statistical software R (R Development Core Team, 2014) and the survey package (Lumley, 2014) to inspect the data. The final models were specified in Mplus 7.4 (Muthén & Muthén, 2012). All models took into consideration survey weights to obtain representative results for students in upper secondary schools in Baden-Württemberg.

In order to report meaningfully interpretable coefficients, we present fully standardized coefficients, meaning that both the dependent and continuous independent variables were standardized. We also present partially standardized coefficients, meaning that only the dependent variable was standardized (also referred to as Cohen's *d*; Cohen, 1988). Continuous variables were centered. The partially standardized coefficients might be especially useful for interpreting effects of dichotomous variables. With regard to the fully standardized solution, the interaction terms were standardized before we included them in the regression models. In order to explore and interpret possible interaction effects, we additionally

[1] Due to the different vocational school types that were considered in the TOSCA studies, we also specified models with dummy-coded variables for every type of vocational school as additional robustness checks. The results did not differ meaningfully.

estimated simple main effects between the two cohorts for young women and men and school types for statistically significant three-way interactions by using the model constraint option in Mplus 7.4. Estimating simple main effects to interpret interactions is also recommended by Jaccard, Wan, and Turrisi (1990). Furthermore, we also calculated structure coefficients (e.g., Courville & Thompson, 2001) to gain further insights into the dynamics of our data. Structure coefficients indicate the proportion of the multiple correlation that can be accounted for by the first-order correlation. When multicollinearity is high, the beta weights might be relatively small. However, structure coefficients are able to indicate this more precisely.

Effect sizes. Regarding the interpretation of effect sizes and on the basis of a literature review, as suggested by Henson (2006), we argue that effect sizes of $d > 0.05$ should be considered practically relevant. As can be seen in the literature, this seems to be the average amount of growth that can be expected from a half to 1 year of schooling (e.g., Hill, Bloom, Black, & Lipsey, 2008; Low, Yoon, Roberts, & Rounds, 2005; Low, 2009; Nagy et al., 2010; Wagner, Rose, Dicke, Neumann, & Trautwein, 2014). However, as stated in Henson (2006), benchmarks should be used cautiously.

Cluster structure. Students from the same class or school cannot be treated as independent observations because they are more similar to each other than they are to students from other classes or schools. Not considering this cluster structure leads to overestimated standard errors (Snijders & Bosker, 2012). To address the clustered data structure (students were nested within classes), standard errors were adjusted by applying a design-based correction as implemented in Mplus (Muthén & Muthén, 2012), which automatically takes the multilevel structure into account and makes use of a sandwich estimator (see, e.g., Asparouhov, 2005; Muthén & Satorra, 1995). Here, we followed McNeish, Stapleton, and Silverman's (2016) recommendations as they pointed out that alternative design-based methods (or population-averaged methods) can be more intuitive and do not rely on assumptions that are inherent in the specification of random effects in hierarchical linear modeling. Design-based methods allow the researcher to adjust the standard errors of estimates and fit statistics for the nested structure of the data and have been shown to perform well in various different nested data settings (e.g., Stapleton, Yang, & Hancock, 2016).

Missing values. Missing values are a common problem in the social sciences, and several approaches have been implemented to account for missing values in a meaningful way (e.g., Enders, 2010; Graham, 2009). There is a growing consensus that approaches such as multiple imputation (MI) or full information maximum likelihood (FIML) estimation are superior to traditional methods (e.g., complete case analysis or pairwise deletion). For all outcomes except math achievement and all independent variables, missing values were addressed with full information maximum likelihood in Mplus 7.4 (Muthén & Muthén, 2012). There were no missing values on the math achievement tests as we used plausible values that were generated for every student and the primary analysis of the TOSCA study (Nagy et al., 2010).

Results

In the first step, the two cohorts were compared with respect to possible differences in the covariates (see Table 1). Overall, these preexisting differences between the two cohorts seemed to be of small practical relevance. Differences were found for age ($d = -0.22$, $p < .001$), largely due to the TOSCA 2002 assessment taking place a little bit later in the school year because of an organizational issue. However, because this difference applied equally to young women and men, it should not have had any effect on the results. Furthermore, we controlled for age in all analyses. In addition, a difference in the number of books available in the home ($d = -0.06$, $p = .021$) was significant, whereas differences on all other variables (including gender) were not significant.

Next, we compared the lengths of time (in hours per week) allocated to mathematics by gender between the two cohorts before and after the reform. Table 2 shows a difference in the average amount of time allocated to math for both young men (3.5 min per week) and young women (19.7 min per week) and an average increase in the total sample after the reform (12.2 min). As expected, the average amount of time allocated to mathematics increased more for young women than for young

Table 1
Descriptive Statistics for the Two Cohorts

Variable	TOSCA 2002	TOSCA 2006	Effect size	p
Gender (% female)	54.0%	53.1%	.98	.679
Immigration background (% immigrants)	20.0%	20.8%	1.08	.237
HISEI	59.16 (15.57)	58.49 (15.73)	−.04	.120
Books	5.64 (1.22)	5.57 (1.23)	−.06	.021
Age	19.56 (.79)	19.40 (.65)	−.22	<.001
Math achievement	50.10 (9.82)	51.07 (9.42)	.10	.002
Math self-concept	2.76 (.81)	2.70 (.85)	−.08	.003
Realistic vocational interests	2.08 (.74)	2.24 (.80)	.20	<.001
Investigative vocational interests	2.60 (.83)	2.64 (.81)	.04	.138

Note. Weighted results. For dichotomous dependent variables, logistic regression was used to test the differences. For continuous dependent variables, linear regression was used. HISEI = highest international socioeconomic index; TOSCA = Transformation of the Secondary School System and Academic Careers. Effect sizes: for dichotomous dependent variables, odds ratios (*OR*s) are displayed; for continuous dependent variables, Cohen's d (Cohen, 1988) is displayed.

Table 2
Time Allocated to Mathematics Before and After the Reform

Group	Enrollees in advanced math courses (TOSCA 2002)	Average allocated time (TOSCA 2002)	Average allocated time (TOSCA 2006)	Increase	*p*
Young men	44.7%	3.92 hr (177 min)	4 hr (180 min)	3.49 min	.007
Young women	27.9%	3.56 hr (160 min)	4 hr (180 min)	19.68 min	<.001
Total	35.5%	3.73 hr (167 min)	4 hr (180 min)	12.17 min	<.001

Note. TOSCA = Transformation of the Secondary School System and Academic Careers. Results for TOSCA 2002 are based on self-reported course choice. The analyses took into consideration the survey weights and clustered structure of the data. One lesson lasted for 45 min. In TOSCA 2006, the average time allocated by young men and women was equal because of the mandatory advanced course.

men as shown by a significant Gender × Cohort interaction ($B = 16.20$, $p < .001$).

Test of Advanced Mathematics Achievement

We hypothesized that the gender difference in math achievement in favor of young men would be smaller after the reform that introduced the obligatory advanced mathematics course for both young men and women. To test our prediction, we used multiple regression analyses to explore a possible difference between the two cohorts in advanced mathematics achievement (see Table 3).

The Cohort × Gender interaction was statistically significant ($d = 0.14$, $p = .025$, 95% CI [0.01, 0.26]). In line with our hypothesis, the interaction indicated a smaller difference between young women and men after the reform than before (see Figure 2). This was mainly due to a higher average level of young women's achievement after the reform ($d = 0.14$, $p = .002$, 95% CI [0.05, 0.22]), whereas young men's achievement did not differ before and after the reform ($d = 0.00$, $p = .988$, 95% CI [−0.11, 0.11]). The Cohort × School Type interaction ($d = 0.08$, $p = .255$, 95% CI [−0.08, 0.22]) was not statistically significant, but the Cohort × Gender × School Type interaction had a significant regression weight ($d = -0.19$, $p = .029$, 95% CI [−0.35, 0.02]), indicating that the effects of the reform differed between the different school types. Our results indicate a three-way interaction between Cohort × Gender × School Type. Exploring this interaction revealed statistically significant differences for young women, but not for young men, before versus after the reform for general gymnasiums but not for vocational gymnasiums, in favor of the cohort that was measured after the reform. However, for young men, the effect of the reform was not statistically significantly different between vocational gymnasiums and general gymnasiums.

Table 3
Predicting Advanced Mathematics Achievement: Results From Multiple Regressions Models

Predictor	B	*p*	*SE*	*d*[a]
Cohort (T2 = 1)	.00	.988	.06	.00
Gender (f = 1)	−.58	<.001	.04	−.58
HISEI	.00	.912	.01	.00
Books	.07	<.001	.01	.06
Immigration background (=1)	−.16	<.001	.03	−.16
Age	−.18	<.001	.02	−.24
School type (VS = 1)	−.61	<.001	.06	−.61
Cohort × Gender	.14	.025	.06	.14
Cohort × School Type	.09	.255	.07	.08
Cohort × Gender × School Type	−.19	.029	.09	−.19
R^2		.23		

Note. All coefficients are fully standardized. Continuous variables are centered. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational school.
[a] The dependent variable is standardized (Cohen, 1988).

Math Self-Concept

With regard to math self-concept, we expected a larger gender difference after the reform. In line with our expectations, and as shown in Table 4, the moderating effect of gender on the relation between cohort and self-concept was statistically significant ($d = -0.16$, $p < .001$, 95% CI [−0.27, −.04]). The larger gender difference after the reform was the result of a statistically significantly lower average math self-concept for young women after the reform ($d = -0.19$, $p < .001$, 95% CI [−0.27, −0.11]) compared with before the reform. For young men, math self-concept did not differ significantly before versus after the reform ($d = 0.04$, $p = .433$, 95% CI [−0.18, 0.08]). The other two interaction effects, Cohort × School Type ($d = -0.03$, $p =$ .619, 95% CI [−0.16, 0.09]) and Cohort × Gender × School Type ($d = 0.11$, $p = .157$, 95% CI [−0.04, 0.27]), were both not statistically significant.

Realistic and Investigative Vocational Interests

According to our hypotheses, we expected larger gender differences in realistic and investigative interests after the reform. As reported in Table 5, we found a significant and negative interaction between cohort and gender in predicting realistic vocational interests ($d = -0.15$, $p = .007$, 95% CI [−0.26, −0.04]), thus indicating a larger gender difference after the reform than before. This larger gender difference resulted from a significantly higher mean score for young men ($d = 0.27$, $p < .001$, 95% [0.19, 0.35]) and a smaller, albeit also significantly higher mean score for young women ($d = 0.12$, $p < .001$, 95% [0.05, 0.19]) after the reform (see Figure 2).

In addition to realistic vocational interests, we tested for a gender difference in investigative interests (see Table 6). Taking a closer look at our results, we found a significant interaction effect (see Figure 2), indicating a larger gender difference in investigative vocational interests after the reform ($d = -0.12$, $p = .019$, 95% CI [−0.23, −0.02]). No significant difference between before and after the reform was found for young women in investigative interests ($d = -0.01$, $p =$ .773, 95% CI [−0.09, 0.07]), but young men showed, on average, a

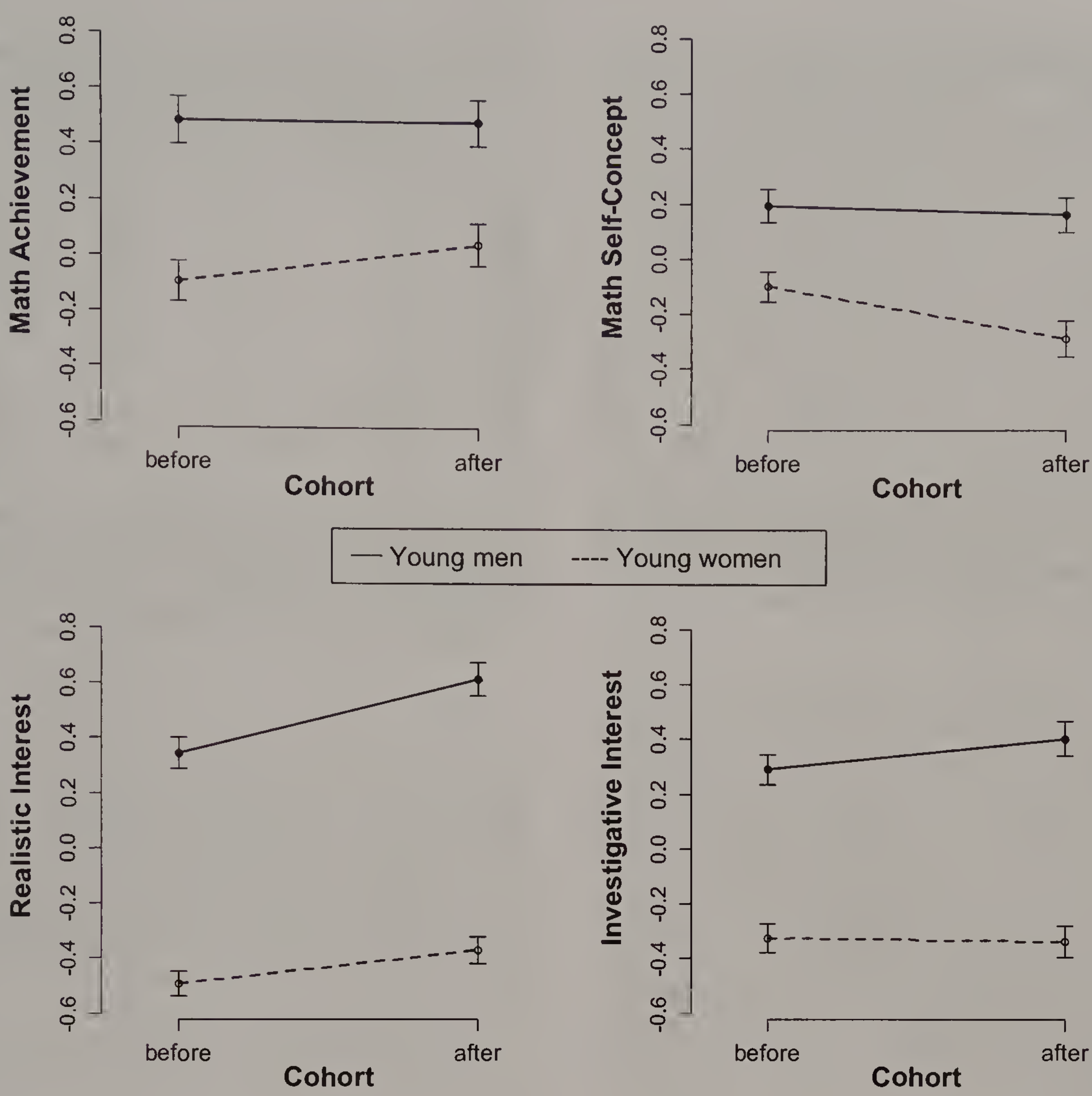


Figure 2. Plots of the moderating effect of gender on the relation between reform and math achievement, math self-concept, realistic interests, and investigative interests with 95% confidence intervals. The dependent variables are presented in standard deviation units. Note that gender differences before the reform were statistically significant for achievement ($d = -0.58$, $p < .001$), math self-concept ($d = -0.29$, $p < .001$), realistic interests ($d = -0.84$, $p < .001$), and investigative interests ($d = -0.62$, $p < .001$). Gender differences after the reform were statistically significant for achievement ($d = -0.44$, $p < .001$), math self-concept ($d = -0.45$, $p < .001$), realistic interests ($d = -0.99$, $p < .001$), and investigative interests ($d = -0.74$, $p < .001$).

higher level of interest after the reform ($d = 0.11$, $p = .01$, 95% CI [0.03, 0.21]. For both outcomes, the Cohort × School Type interaction and the Cohort × Gender × School Type interaction were not statistically significant (see Table 6).

The results for the multivariate approach were similar to the results for the univariate approach: The Wald test for the interaction effect was statistically significant, $\chi^2(12) = 55.06$, $p < .001$. Furthermore, even after the Benjamini-Hochberg corrections, all interaction effects remained statistically significant in the multivariate and univariate approaches. Overall, we found that the structure coefficients supported our results regarding multiple linear regression models and the interpretation of the relevance of the Cohort × Gender interaction for all outcome variables (see Table 7).

Field of Study at University

Whether or not the upper secondary school reform had an effect on university subject choices was handled as an open research question. Therefore, we did not formulate an explicit hypothesis with regard to this construct. The results presented here are based on an analysis that considered only students who did not intend to become teachers.[2] As reported in Table 8, none of the additional interaction effects were statistically significant. Thus, a potential shift, which would go along with an increase in women enrolling in STEM

[2] The pattern of gender differences in the literature varies with respect to different professions within the STEM fields. Whereas a larger percentage of young men than women tend to choose mathematically intensive STEM subjects, gender differences are much less pronounced with regard to STEM teaching professions (Watt, Richardson, & Devos, 2013). To meet this objective, we excluded teaching students from our analysis. However, robustness checks did not reveal any substantial difference between the results of these two groups of students. Furthermore, although men tended to start their studies a bit later (e.g., due to mandatory community or military services), we did not find significant gender differences before and after the reform regarding students who attended university and those who did not.

Table 4
Predicting Advanced Math Self-Concept: Results From Multiple Regressions Models

Predictor	B	p	SE	d^a
Cohort (T2 = 1)	−.04	.433	.04	−.04
Gender (f = 1)	−.29	<.001	.03	−.29
HISEI	.01	.370	.01	.00
Books	.05	<.001	.01	.04
Immigration background (=1)	.00	.925	.03	.00
Age	−.13	<.001	.02	−.18
School type (VS = 1)	.06	.131	.04	.06
Cohort × Gender	−.16	<.001	.06	−.16
Cohort × School Type	−.03	.619	.06	−.03
Cohort × Gender × School Type	.11	.157	.08	.11
R^2		.05		

Note. All coefficients are fully standardized. Continuous variables are centered. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school.
[a] The dependent variable is standardized (Cohen, 1988).

Table 6
Predicting Investigative Vocational Interests: Results From Multiple Regressions Models

Predictor	B	p	SE	d^a
Cohort (T2 = 1)	.11	.01	.04	.11
Gender (f = 1)	−.62	<.001	.03	−.62
HISEI	.00	.745	.01	.00
Books	.11	<.001	.01	.09
Immigration background (=1)	.01	.668	.03	.01
Age	−.03	.045	.01	−.04
School type (VS = 1)	.07	.142	.05	.07
Cohort × Gender	−.12	.019	.05	−.12
Cohort × School Type	−.05	.371	.06	−.05
Cohort × Gender × School Type	.02	.770	.08	.02
R^2		.12		

Note. All coefficients are fully standardized. Continuous variables are centered. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school; TOSCA = Transformation of the Secondary School System and Academic Careers.
[a] The dependent variable is standardized (Cohen, 1988).

subjects at university was not found in our data set (Cohort × Gender: OR = 1.02, p = .838, 95% CI [0.86, 1.21]). We further tested for potential differences between students who provided information about their university subject and those who did not. Results revealed that women (OR = 0.73, $p < .001$) and students from vocational schools (OR = 0.54, $p < .001$) as well as older students ($B = -.20$, $p < .001$) were less likely to report their subject, whereas students with a higher HISEI ($B = .28$, $p < .001$), more books at home (B = .33, $p < .001$), and higher cognitive abilities ($B = .28$, $p < .001$) reported their subject more often. We controlled for these variables in all analyses. It is important to note that these differences did not differ significantly between the two cohorts, as shown by the Wald test, $\chi^2(7) = 7.75$, $p = .356$.

Discussion

In the current study, we examined effects of a higher secondary school education reform on STEM-related outcomes in a large and representative sample. The reform is of high theoretical and practical interest because it abolished a prior imbalance between young men and women in taking advanced math courses. High school coursework in math has been shown to be related to STEM career choices as well as to math achievement, math self-concept, and vocational interests, all of which are important predictors of STEM career choices. Therefore, we expected that the effects of the reform on these outcomes would differ by gender. Overall, the results supported most of our predictions. First, there were significant gender differences in all outcomes before the reform, with higher scores for young men than for young women. Second, we found differential effects of the reform for young women and men in all outcomes except field of study at university. However, the direction of the effects differed: The gender difference in math achievement was smaller after the reform, but gender differences in math self-concept and STEM-related vocational interests were even larger after the reform than before. However, the larger gender difference after the reform in math self-concept was based

Table 5
Predicting Realistic Vocational Interests: Results From Multiple Regressions Models

Predictor	B	p	SE	d^a
Cohort (T2 = 1)	.27	<.001	.04	.27
Gender (f = 1)	−.84	<.001	.03	−.84
HISEI	−.04	.004	.01	.00
Books	.04	.001	.01	.03
Immigration background (=1)	−.08	.002	.03	−.08
Age	−.03	.013	.01	−.05
School type (VS = 1)	.09	.099	.06	.09
Cohort × Gender	−.15	.007	.06	−.15
Cohort × School Type	.00	.932	.07	.01
Cohort × Gender × School Type	.00	.948	.09	.01
R^2		.22		

Note. All coefficients are fully standardized. Continuous variables are centered. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school.
[a] The dependent variable is standardized (Cohen, 1988).

Table 7
Structure Coefficients for Multiple Linear Regression Models

Predictor	Advanced mathematics	Math self-concept	Realistic interests	Investigative interests
Cohort (T2 = 1)	.11	−.18	.21	.05
Gender (f = 1)	−.54	−.74	−.96	−.95
HISEI	.28	.22	−.00	.15
Books	.32	.25	.03	.28
Immigration background (=1)	−.25	−.14	−.09	−.07
Age	−.49	−.50	−.02	−.03
School type (VS = 1)	−.70	−.02	.08	.00
Cohort × Gender	−.23	−.61	−.47	−.55
Cohort × School Type	−.42	−.07	.14	.00
Cohort × Gender × School Type	−.47	−.25	−.20	−.28

Note. The table displays structure coefficients (e.g., Courville & Thompson, 2001) for each predictor of all four multiple linear regression models. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school; TOSCA = Transformation of the Secondary School System and Academic Careers.

Table 8
Predicting Field of Study at University: Results From Multiple Logistic Regressions Models

Predictor	*OR*	CI		*p*
Cohort (T2 = 1)	.97	.85	1.11	.702
Gender (f = 1)	.37	.32	.42	<.001
HISEI	.90	.82	.98	.022
Books	.90	.82	.99	.037
Immigration background (=1)	.97	.88	1.07	.538
Age	.83	.75	.92	<.001
School type (VS = 1)	1.08	.90	1.31	.411
Cohort × Gender	1.02	.86	1.21	.838
Cohort × School Type	1.01	.86	1.20	.871
Cohort × Gender × School Type	1.01	.87	1.16	.948
Pseudo-R^2		.24		

Note. The table displays standardized results where mathematics, engineering, computer science, and physics were coded as STEM subjects. Odds ratios significantly larger than 1 indicate a higher likelihood of studying STEM subjects. T2 = TOSCA 2006; HISEI = highest international socioeconomic index; VS = vocational upper secondary school; TOSCA = Transformation of the Secondary School System and Academic Careers.

on young women's lower scores, whereas young men's scores did not differ. Also, the greater differences in vocational interests were due to young men's higher interests after the reform, whereas young women's interests were only slightly higher (realistic) or did not differ (investigative). Third, we found no overall effect of the reform on gender differences in the choice of STEM subjects at university.

Differential Reform Effects for Young Men and Women

The effects of the reform on math achievement are in accordance with previous research that reported positive effects of course level on achievement, which can be attributed to more demanding curricula, more teaching time, and larger weights from grades in advanced courses with respect to their contribution to final GPA (e.g., Brunello & Checchi, 2007; Gamoran & Mare, 1989; Hanushek & Wössmann, 2006; Kelly, 2004; Lucas, 2001). Presumably because a larger proportion of young men than women had chosen advanced courses in math before the reform, but all students took the same math course after the reform, young women were able to come closer to young men's math achievement, although there was still a significant gender difference after the reform. In addition, there was a difference in teaching time between the courses before versus after the reform, with more lessons taught per week in the advanced course (five lessons) than in the basic course (three lessons). Although meta-analyses do not suggest a clear pattern with regard to the effects of extended learning time on achievement, most studies have shown zero to small positive effects (e.g., Patall, Cooper, & Allen, 2010; Scheerens & Hendriks, 2014). Thus, the difference in teaching time might provide a possible explanation for the differential effects of the reform on young women's and men's math achievement. However, we cannot explicitly test for or disentangle the effects of instructional time or course level on our results at this point.

Against this background, the larger gender differences after the reform with respect to math self-concept and STEM-related interests might come as a surprise at first glance. A change in reference group provides a good explanation for the larger gender difference in math self-concept after the reform: It is a common finding that social comparisons are central for the development of students' self-concept. In evaluating their own abilities, individuals refer not only to their own prior achievement in a domain, but also to the level of achievement they perceive in their surroundings (e.g., Marsh, 2005; Niepel, Brunner, & Preckel, 2014; Trautwein et al., 2006). As discussed above, students' achievement differs between advanced and basic courses; thus, both courses provide different frames of reference for social comparisons. Higher course levels are usually associated with negative effects on students' evaluations of their own abilities after individual ability is controlled for (Marsh, 2005; Trautwein et al., 2006). Before the reform, young women tended to choose basic courses in math where they were surrounded by (on an average) a weaker reference group, compared with students in advanced courses. Therefore, they perceived their own math ability in comparison with other, on average, lower achieving classmates. After the reform, all students were instructed at the same course level. Consequently, after the reform, young women could compare their own achievement with the achievement of all other students in their class, which included students with relatively lower achievement but also those with relatively higher achievement. It is therefore likely that the reason why young women's evaluation of their own math abilities was somewhat lower was due to the, on average, higher achieving reference group. There was no significant difference in young men's self-concept after the reform, which can be explained by the proportions of young men in advanced and basic courses before the reform, as they participated in advanced and basic courses in almost equal parts before the reform. According to the literature described above, it is therefore likely that possible effects of course level on young men's self-concept cancelled each other out. These explanations are further supported by the fact that young women's math self-concept in basic courses before the reform was statistically lower, compared with young women's self-concept after the reform ($d = -0.12, p < .001$), whereas the reverse was true for young women in advanced courses before the reform ($d = 0.83$, $p < .001$). Furthermore, the difference between young men and women in basic courses was not statistically significant ($d = -0.07$, $p = .086$), whereas the gender gap for advanced course students was statistically significant, favoring young men ($d = -0.13$, $p = .001$).

In our study, we found larger gender difference after the reform in realistic and investigative interests as well, but in contrast to math self-concept, the greater differences were based on young men's higher levels of interests after the reform, whereas young women showed only slightly higher interests (realistic) or even similar scores (investigative) after the reform. There is a gap in research on how vocational interests might be related to course level. However, as reported in the Introduction, previous research has indicated positive relations between individual levels of math achievement and realistic and investigative interests (Ackerman & Heggestad, 1997; Anthoney & Armstrong, 2010). Furthermore, previous research has shown negative effects of the mean level of

achievement on domain-specific levels of interest (Köller, Trautwein, Lüdtke, & Baumert, 2006; Schurtz, Pfost, Nagengast, & Artelt, 2014; Trautwein et al., 2006) and initial findings with respect to vocational interests. These findings indicate that there might be positive effects of the mean level of math achievement on realistic and investigative vocational interests (Cambria et al., 2016). However, as these findings provide only initial indications on how vocational interests might be related to class level, they enable us to discuss our findings only on a speculative basis. Thereby, one could argue that there might be a positive association between class level and students' realistic and investigative interests, but this association differs by gender, with larger associations for young men than for young women. Previous research on vocational interests has indeed indicated differential associations between ability and vocational interests, although such findings have so far been limited to general cognitive ability and have not been applied to math (e.g., Reeve & Heggestad, 2004). However, more research is needed to explore the relation between class level and vocational interests for young women as well as for young men.

Although we found differential effects of the reform on central predictors of STEM career choices, we found no difference in gender ratios in the numbers of students who chose to study STEM university subjects. There are two aspects to consider when interpreting the absence of effects of the reform on gender differences in STEM university subject choices. First, we found opposite effects of the reform on gender differences in four important predictors of STEM career choices: Whereas differences in math achievement were eliminated, differences in math self-concept and both interest facets were larger. Consequently, it is possible that the effects of the reform on the predictors cancelled each other out, with the consequence that no effect on the choice of STEM subjects remained. Second, choosing a university subject is a complex process that involves numerous factors (see Schoon & Eccles, 2014). The reform influenced students' upper secondary high school coursework, but it did not directly affect other structural factors or the wider context they grew up in, such as their family structure, the role models they perceived, or their stereotypical views of STEM professions.

Practical Implications

Our study adds to the increasing number of studies that have found intended as well as unintended effects of educational reforms. In fact, educational policy reforms do not necessarily improve educational outcomes but can instead result in numerous unintended consequences. In addition, the aspects of the reforms most likely interact differently with different student characteristics, even if such aspects are well-structured and carefully planned (Gross, Booker, & Goldhaber, 2009). For instance, studies by Gross et al. (2009); Domina, McEachin, Penner, and Penner (2015), and Lee and Reeves (2012) showed that school reforms could have differential effects for minority students (e.g., African American and Hispanic students) or could vary for specific school districts. The results show that school reforms can have differential effects on several outcomes, and such outcomes can even differ for particular subgroups such as young women and men; not every well-intentioned reform will reach all goals, and some might even backfire.

Unintended consequences of reforms can be attributed to, among other factors, the complex nature of establishing and especially of implementing reforms (e.g., McLaughlin, 1987; Young & Lewis, 2015) in the education sector as a "loosely coupled system" (Porter, Fusarelli, & Fusarelli, 2015, p. 114). Conversely, with regard to the current study, one might argue that the higher achievement and realistic interests that came with this reform came at a price—a lower math self-concept for young women—which had to be expected given the change in reference group.

Although high school coursework is central to young people's career choices, and although we found differential effects of the reform on central predictors of STEM career choices for young men and women, we did not find effects of the reform on gender differences in the choice of STEM university subjects, which indicates that one single reform might not significantly influence students' career choices. In the complex context that young people grow up in, there is a cumulative process of multiple experiences that shape young people's academic attitudes and behavior, such as career choice (cf. Schoon & Eccles, 2014). Influencing gender differences in high school course selection by restricting choice options might be one way to balance some gender differences in the STEM context, namely, gender differences in math achievement. However, reforming course choice options does not necessarily impact any of the reasons for why young women are less likely to choose advanced math courses than young men (e.g., gender stereotypes, different expectancies of parents, teachers, peers; cf. Schoon & Eccles, 2014; Wigfield & Eccles, 2000). Such high school reforms might therefore be "too little too late" to increase gender equity in the STEM fields in a meaningful and sustainable way. Furthermore, although course-taking gaps in other countries have narrowed in recent decades (e.g., Domina & Saldana, 2012; Osborne & Dillon, 2008), subsequent changes in STEM career plans do not seem to be of considerable size (Jerrim & Schoon, 2014).

Limitations and Further Research

The current study demonstrates that intensifying school curricula and providing equal access to advanced courses "does not necessarily level the [educational] playing field" with regard to all important outcomes (Domina & Saldana, 2012, p. 688). Although our investigation was based on a strong data set, some limitations should be kept in mind when interpreting the results. First, our results were limited to the domain of math. Math is a key domain within the STEM fields (Ma & Johnson, 2008; Sells, 1980), and math achievement, self-concept, and interests are very important for math-intensive STEM career choices (e.g., Parker et al., 2012; Schoon & Eccles, 2014). Nevertheless, other STEM domains such as physics or chemistry are also meaningful for later math-intensive STEM career choices (e.g., Hazari, Sonnert, Sadler, & Shanahan, 2010), and gender differences in such high school courses are often even larger than in math (e.g., NSF, 2015). Evaluating the effects of a reform on central STEM outcomes in these domains might therefore provide additional information about effects on important predictors of math-intensive STEM career choices.

Second, the current study was based on cross-sectional data. According to Shadish, Cook, and Campbell (2002), quasi-experiments lack "random assignment of units to conditions" (p. 104), which may lead to selection bias. We attempted to address these challenges by using a lagged cohort control design that should have led to relatively small selection differences between cohorts (drawn from the same schools). We additionally checked for potential differences between cohorts and used covariates to control for these.

Third, besides these methodological issues, there are other possible reasons for the results that we found. Our results may be explained by the multidimensional structure of the reform. As stated by Malen and Knapp (1997), "policy takes many forms, performs many functions, and begets many effects," which is why "it is difficult to get a fix on the boundaries, let alone the 'workings' of a policy or a set of policies" (p. 419). In our case, as mentioned, not only did time vary between the groups before and after the reform, but the reference groups and course levels also varied. Therefore, the effects of the reform cannot be directly attributed to one specific aspect or mechanism of the reform in a causal manner but must be interpreted from within the multilayered framework of the entire policy reform.

However, as society is constantly changing, it would be reasonable to expect main and interaction effects that indicate the increased participation of young women in STEM classes because they are now as able to do so as young men. However, the results of our study instead indicate the opposite pattern. Regarding this point, it is also important to mention that society's growing interest and all resulting efforts had already increased in the beginning of this century and not just between these two cohorts in particular (Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung, 2002; NSF, 2000). In addition, we checked closely whether any other educational reforms had been implemented between the two cohorts, but this was not the case.

Further research should address the question of whether effects, such as the drop in self-concept, can be found in different subsamples. This refers to questions such as whether such effects can be found for all young women or only the subsample of those who would have chosen basic courses if they had been allowed to, and whether similar effects can be found for young men who would have chosen basic courses if they had been allowed to.

Fourth, our results are limited to the issue of gender differences in STEM career choices at the end of secondary education, and more research is needed to explore the complex pattern of gender differences in the STEM fields throughout students' educational careers. In our study, we focused on important predictors of STEM career choices as well as students' choice of university major in the STEM fields. Therefore, our results provide insights into various effects of the reform on central STEM outcomes. However, regarding the issue of gender differences in the STEM area, not only do women tend to choose such majors less frequently than their male counterparts, but women also drop out of university at higher rates (Ackerman, Kanfer, & Beier, 2013; Perez-Felkner, McDonald, & Schneider, 2014). Considering social comparison processes, one could possibly argue that women entering the STEM fields are likely to experience such comparison processes during their studies, where they need to deal with other high-achieving students. Experiencing such comparison processes at an earlier point in high school might therefore make women less likely to pursue such careers and—consequently—less vulnerable to dropping out of STEM fields during college.

Furthermore, prior work on the development of interest suggests that interest takes time to develop (see Hidi & Renninger, 2006) and that such a change in upper secondary high school coursework as investigated in the present study might be less related to students' vocational interests than to their achievement and self-concept or that such effects might take longer. In this study, we investigated effects of changes in coursework requirements on students' interests 1.5 years after they started taking these high school courses. It might be the case that such a time period is insufficient to fully study effects on interest developments and that effects would be different or more pronounced if more time could have elapsed between when the students began taking these high school courses and the measurement point. Further research spanning a longer time frame is needed to test such propositions as well as to develop more potent remedies for the gender differences that still exist.

Conclusion

The present study was aimed at taking a closer look at effects of high school coursework on gender differences in math-intensive STEM fields. To this end, we investigated effects of a statewide educational reform in Germany with a large representative sample. The reform required all students to take advanced courses in math and eliminated the prior imbalance between young men and women in choosing such courses. Our results showed that it is crucial to take multiple aspects into consideration in order to obtain insights into possible differential effects of changes in coursework requirements. Although requiring all students to take advanced math courses appears to be adequate for reducing gender differences in math achievement, it seems that young women were not aware of this: Young men and women's achievement differed less after the reform, but young women showed an even lower self-concept compared with young men than had been there before the reform. With respect to realistic and investigative interests, although young women showed no or only slightly higher interests after the reform, the interests of young men were substantially higher after the reform. Mechanisms that ensure that all students will benefit in comparable ways from such school reforms and impede negative side effects, such as those found for young women's self-concept, should be a primary focus of future research.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121,* 219–245. http://dx.doi.org/10.1037/0033-2909.121.2.219

Ackerman, P. L., Kanfer, R., & Beier, M. E. (2013). Trait complex, cognitive ability, and domain knowledge predictors of baccalaureate success, STEM persistence, and gender differences. *Journal of Educational Psychology, 105,* 911–927. http://dx.doi.org/10.1037/a0032338

Alarcon, G. M., & Edwards, J. M. (2013). Ability and motivation: Assessing individual factors that contribute to university retention. *Journal of Educational Psychology, 105,* 129–137. http://dx.doi.org/10.1037/a0028496

Anthoney, S. F., & Armstrong, P. I. (2010). Individuals and environments: Linking ability and skill ratings with interests. *Journal of Counseling Psychology, 57,* 36–51. http://dx.doi.org/10.1037/a0018067

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12,* 411–434. http://dx.doi.org/10.1207/s15328007sem1203_4

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B: Methodological, 57,* 289–300.

Bergmann, C., & Eder, F. (2005). *AIST-R: Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R)–Revision* [General Interest Structure Test and Environmental Structure Test–Revision]. Göttingen, Germany: Beltz.

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review, 15,* 1–40. http://dx.doi.org/10.1023/A:1021302408382

Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy, 22,* 782–861. http://dx.doi.org/10.1111/j.1468-0327.2007.00189.x

Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung. (2002). *Frauen in den ingenieur- und naturwissenschaftlichen Studiengängen* [Women in engineer and natural science degree programs]. (Bericht der BLK, vol. 2). Retrieved from http://www.blk-bonn.de/papers/heft100.pdf

Cambria, J., Brandt, H., & Nagengast, B., & Trautwein, U. (2016). *Vocational interests: The impact of class achievement and gender.* Manuscript in preparation.

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin, 135,* 218–261. http://dx.doi.org/10.1037/a0014412

Chen, S.-K., Yeh, Y.-C., Hwang, F.-M., & Lin, S. S. J. (2013). The relationship between academic self-concept and achievement: A multicohort–multioccasion study. *Learning and Individual Differences, 23,* 172–178. http://dx.doi.org/10.1016/j.lindif.2012.07.021

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students' mathematics self-concept. *American Educational Research Journal, 50,* 925–957. http://dx.doi.org/10.3102/0002831213489843

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: B is not enough. *Educational and Psychological Measurement, 61,* 229–248.

Denissen, J. J. A., Zarrett, N. R., & Eccles, J. S. (2007). I like to do it, I'm able, and I know I am: Longitudinal couplings between domain-specific achievement, self-concept, and interest. *Child Development, 78,* 430–447. http://dx.doi.org/10.1111/j.1467-8624.2007.01007.x

Domina, T., McEachin, A., Penner, A., & Penner, E. (2015). Aiming high and falling short: California's eighth-grade algebra-for-all effort. *Educational Evaluation and Policy Analysis, 37,* 275–295. http://dx.doi.org/10.3102/0162373714543685

Domina, T., & Saldana, J. (2012). Does raising the bar level the playing field?: Mathematics curricular intensification and inequality in American high schools, 1982–2004. *American Educational Research Journal, 49,* 685–708. http://dx.doi.org/10.3102/0002831211426347

Eccles, J. S., Adler, T., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. C. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–121). San Francisco, CA: Freeman & Co.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136,* 103–127. http://dx.doi.org/10.1037/a0018053

Enders, C. K. (2010). *Applied missing data analysis. Methodology in the social sciences.* New York, NY: Guilford Press.

Evans, M. D. R., Kelley, J., Sikora, J., & Treiman, D. J. (2010). Family scholarly culture and educational success: Books and schooling in 27 nations. *Research in Social Stratification and Mobility, 28,* 171–197. http://dx.doi.org/10.1016/j.rssm.2010.01.002

Federal Statistical Office. (Ed.), (2008). *Bildung und Kultur—Studierende an Hochschulen Wintersemester 2007/2008* (Fachserie 11, Reihe 4.1) [Education and Culture - Students at Universities in Wintersemester 2007/2008]. Wiesbaden, Germany: Statistisches Bundesamt.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? *American Journal of Sociology, 94,* 1146–1183. http://dx.doi.org/10.1086/229114

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21,* 1–56. http://dx.doi.org/10.1016/0049-089X(92)90017-B

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social Science Research, 25,* 201–239. http://dx.doi.org/10.1006/ssre.1996.0010

Gottfredson, L. S. (1986). Occupational aptitude patterns map: Development and implications for a theory of job aptitude requirements. *Journal of Vocational Behavior, 29,* 254–291. http://dx.doi.org/10.1016/0001-8791(86)90008-4

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Gross, B., Booker, T. K., & Goldhaber, D. (2009). Boosting student achievement: The effect of comprehensive school reform on student achievement. *Educational Evaluation and Policy Analysis, 31,* 111–126. http://dx.doi.org/10.3102/0162373709333886

Hanushek, E. A., & Wössmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal, 116*(510), C63–C76. http://dx.doi.org/10.1111/j.1468-0297.2006.01076.x

Hazari, Z., Sonnert, G., Sadler, P. M., & Shanahan, M.-C. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching, 47,* 978–1003.

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist, 34,* 601–629. http://dx.doi.org/10.1177/0011000005283558

Hidi, S., & Ainley, M. (2002). Interest and Adolescence. In F. Pajares & T. C. Urdan (Eds.), *Academic motivation of adolescence* (pp. 247–275). Greenwich, CT: Information Age Publishing.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41,* 111–127. http://dx.doi.org/10.1207/s15326985ep4102_4

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2,* 172–177. http://dx.doi.org/10.1111/j.1750-8606.2008.00061.x

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6,* 35–45. http://dx.doi.org/10.1037/h0040767

Holland, J. L. (1966). *The psychology of vocational choice: A theory of personality types and model environments.* Waltham, MA: Blaisdell.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Huang, J. L., & Pearce, M. (2013). The other side of the coin: Vocational interests, interest differentiation and annual income at the occupation

level of analysis. *Journal of Vocational Behavior, 83,* 315–326. http://dx.doi.org/10.1016/j.jvb.2013.06.003

Humphreys, L. G., & Yao, G. (2002). Prediction of graduate major from cognitive and self-report test scores obtained during the high school years. *Psychological Reports, 90,* 3–30. http://dx.doi.org/10.2466/pr0.2002.90.1.3

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107,* 139–155. http://dx.doi.org/10.1037/0033-2909.107.2.139

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect. *Psychology of Women Quarterly, 14,* 299–324. http://dx.doi.org/10.1111/j.1471-6402.1990.tb00022.x

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321,* 494–495. http://dx.doi.org/10.1126/science.1160364

Jaccard, J., Wan, C. K., & Turrisi, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research, 25,* 467–478. http://dx.doi.org/10.1207/s15327906mbr2504_4

Jansen, M., Lüdtke, O., & Schroeders, U. (2016). Evidence for a positive relation between interest and achievement: Examining between-person and within-person variation in five domains. *Contemporary Educational Psychology, 46,* 116–127. https://doi.org/10.1016/j.cedpsych.2016.05.004

Jansen, M., Scherer, R., & Schroeders, U. (2015). Students' self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology, 41,* 13–24. http://dx.doi.org/10.1016/j.cedpsych.2014.11.002

Jerrim, J., & Schoon, I. (2014). Do teenagers want to become scientists? A comparison of gender differences in attitudes toward science, career expectations, and academic skills across 29 countries. In I. Schoon & J. S. Eccles (Eds.), *Gender differences in aspirations and attainment* (pp. 203–223). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139128933.014

Kelly, S. (2004). Are teachers tracked? On what basis and with what consequences. *Social Psychology of Education, 7,* 55–72. http://dx.doi.org/10.1023/B:SPOE.0000010673.78910.f1

Köller, O., Baumert, J., & Schnabel, K. U. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education, 32,* 448–470. http://dx.doi.org/10.2307/749801

Köller, O., Trautwein, U., Lüdtke, O., & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe [On the Interplay of Academic Achievement, Self-Concept, and Interest in Upper Secondary Schools]. *Zeitschrift fur Pädagogische Psychologie, 20,* 27–39. http://dx.doi.org/10.1024/1010-0652.20.12.27

Köller, O., Watermann, R., Trautwein, U., & Lüdtke, O. (2004). *Wege zur Hochschulreife in Baden-Württemberg: TOSCA–eine Untersuchung an allgemein bildenden und beruflichen Gymnasien* [Ways to qualification for university entrance. TOSCA – A study at traditional and vocational Gymnasium schools]. Opladen, Germany: Leske + Budrich.

Krapp, A. (1999). Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education, 14,* 23–40. http://dx.doi.org/10.1007/BF03173109

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990–2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis, 34,* 209–231. http://dx.doi.org/10.3102/0162373711431604

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45,* 79–122. http://dx.doi.org/10.1006/jvbe.1994.1027

Lippa, R. (1998). Gender-related individual differences and the structure of vocational interests: The importance of the people-things dimension. *Journal of Personality and Social Psychology, 74,* 996–1009. http://dx.doi.org/10.1037/0022-3514.74.4.996

Low, K. S. (2009). *Patterns of mean-level changes in vocational interests: A quantitative review of longitudinal studies* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Champaign, IL.

Low, K. S., Yoon, M., Roberts, B. W., & Rounds, J. (2005). The stability of vocational interests from early adolescence to middle adulthood: A quantitative review of longitudinal studies. *Psychological Bulletin, 131,* 713–737. http://dx.doi.org/10.1037/0033-2909.131.5.713

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science, 1,* 316–345. http://dx.doi.org/10.1111/j.1745-6916.2006.00019.x

Lucas, S. R. (2001). Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology, 106,* 1642–1690. http://dx.doi.org/10.1086/321300

Lumley, T. (2014). *Survey: Analysis of complex survey samples* (R package, version 3.30). Retrieved from https://cran.r-project.org/web/packages/survey/survey.pdf

Ma, X., & Johnson, W. (2008). Mathematics as the critical filter: Curricular effects on gendered career choices. In H. M. G. Watt & J. S. Eccles (Eds.), *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences* (pp. 55–83). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/11706-002

Malen, B., & Knapp, M. (1997). Rethinking the multiple perspectives approach to education policy analysis: Implications for policy-practice connections. *Journal of Education Policy, 12,* 419–445. http://dx.doi.org/10.1080/0268093970120509

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23,* 129–149. http://dx.doi.org/10.3102/00028312023001129

Marsh, H. W. (1992). *Self -Description Questionnaire (SDQ) III: A theoretical and empirical basis for the measurement of multiple dimensions of late adolescent self-concept: A test manual and a research monograph.* Sydney, New South Wales, Australia: Macarthur.

Marsh, H. W. (2005). Big-fish-little-pond effect on academic self-concept. *Zeitschrift fur Padagogische Psychologie/German Journal of Educational Psychology, 19,* 119–129. http://dx.doi.org/10.1024/1010-0652.19.3.119

Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology.* Leicester, UK: British Psychological Society.

Marsh, H. W., Abduljabbar, A. S., Morin, A. J. S., Parker, P. D., Abdelfattah, F., Nagengast, B., & Abu-Hilal, M. M. (2015). The big-fish-little-pond effect: Generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries. *Journal of Educational Psychology, 107,* 258–271. http://dx.doi.org/10.1037/a0037485

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., Nagengast, B., . . . Abu-Hilal, M. M. (2015). The internal/external frame of reference model of self-concept and achievement relations: Age-cohort and cross-cultural differences. *American Educational Research Journal, 52,* 168–202. http://dx.doi.org/10.3102/0002831214549453

Marsh, H. W., & O'Neill, R. (1984). Self-Description Questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement, 21,* 153–174. http://dx.doi.org/10.1111/j.1745-3984.1984.tb00227.x

Marsh, H. W., & Shavelson, R. J. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20,* 107–123. http://dx.doi.org/10.1207/s15326985ep2003_1

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal, 44,* 631–669. http://dx.doi.org/10.3102/0002831207306728

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76,* 397–416. http://dx.doi.org/10.1111/j.1467-8624.2005.00853.x

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal, 35,* 705–738. http://dx.doi.org/10.3102/00028312035004705

McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis, 9,* 171–178. http://dx.doi.org/10.3102/01623737009002171

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2016). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods.* Advance online publication. http://dx.doi.org/10.1037/met0000078

Ministry of Education, Cultural Affairs, Youth, & Sport, Baden-Württemberg. (2002). *Die neue gymnasiale Oberstufe in Baden-Württemberg* (Infodienst schule spezial) [The new upper secondary school in Baden-Württemberg]. Stuttgart, Germany: Ministerium für Kultus Jugend und Sport Baden-Württemberg.

Mullis, I. V. S., Martin, M. O., Beaton, A., Gonzalez, E., Kelly, D., & Smith, T. (1998). *Mathematics and science achievement in the final year of secondary school: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Muthén, B. O., & Muthén, L. K. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267–316. http://dx.doi.org/10.2307/271070

Nagy, G., Garrett, J., Trautwein, U., Cortina, K. S., Baumert, J., & Eccles, J. S. (2008). Gendered high school course selection as a precursor of gendered careers: The mediating role of self-concept and intrinsic value. In J. S. Eccles & H. M. G. Watt (Eds.), *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences* (pp. 115–143). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/11706-004

Nagy, G., & Husemann, N. (2010). Berufliche Interessen vor und nach dem Übergang in die gymnasiale Oberstufe [Vocational interests before and after the transition to Gymnasium upper secondary school]. In W. Bos, E. Klieme, & O. Köller (Eds.), *Schulische Lerngelegenheiten und Kompetenzentwicklung* [Learning opportunities and development of competencies at school]. *Festschrift für Jürgen Baumert* (pp. 59–84). Berlin, Germany: Waxmann.

Nagy, G., Neumann, M., Trautwein, U., & Lüdtke, O. (2010). Voruniversitäre Mathematikleistungen vor und nach der Neuordnung der gymnasialen Oberstufe in Baden-Württemberg [Advanced math before and after the rearrangement of upper secondary school in Baden-Württemberg]. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke, & K. Maaz (Eds.), *Schulleistungen von Abiturienten. Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand* [School achivement of upper secondary school students. The rearrangend upper secondary school on the trial] (pp. 147–180). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften. http://dx.doi.org/10.1007/978-3-531-92037-5_6

Nagy, G., Watt, H. M. G., Eccles, J. S., Trautwein, U., Lüdtke, O., & Baumert, J. (2010). The development of students' mathematics self-concept in relation to gender: Different countries, different trajectories? *Journal of Research on Adolescence, 20,* 482–506. http://dx.doi.org/10.1111/j.1532-7795.2010.00644.x

Niepel, C., Brunner, M., & Preckel, F. (2014). The longitudinal interplay of students' academic self-concepts and achievements within and across domains: Replicating and extending the reciprocal internal/external frame of reference model. *Journal of Educational Psychology, 106,* 1170–1191. http://dx.doi.org/10.1037/a0036307

National Science Foundation (NSF). (2000). *Summary report on the Impact Study of the National Science Foundation's Program for Women and Girls.* Retrieved from https://www.nsf.gov/pubs/2001/nsf0127/nsf0127.pdf

National Science Foundation (NSF). (2013). *Women, minorities, and persons with disabilities in science and engineering.* Retrieved from http://www.nsf.gov/statistics/wmpd/2013/pdf/nsf13304_full.pdf

National Science Foundation (NSF). (2015). *Women, minorities, and persons with disabilities in science and engineering: 2015 Digest* (Special report NSF). Retrieved from https://www.nsf.gov/statistics/2015/nsf15311/digest/

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science, 7,* 384–403. http://dx.doi.org/10.1177/1745691612449021

OECD. (2010). *OECD information technology outlook 2010.* Pisa, Italy: OECD Publishing.

Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections* (A report to the Nuffield Foundation). Retrieved from http://efepereth.wdfiles.com/local-files/science-education/Sci_Ed_in_Europe_Report_Final.pdf

Parker, P. D., Schoon, I., Tsai, Y.-M., Nagy, G., Trautwein, U., & Eccles, J. S. (2012). Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: A multicontext study. *Developmental Psychology, 48,* 1629–1642. http://dx.doi.org/10.1037/a0029167

Pässler, K., Beinicke, A., & Hell, B. (2014). Gender-related differential validity and differential prediction in interest inventories. *Journal of Career Assessment, 22,* 138–152. http://dx.doi.org/10.1177/1069072713492934

Patall, E. A., Cooper, H., & Allen, A. B. (2010). Extending the school day or school year: A systematic review of research (1985–2009). *Review of Educational Research, 80,* 401–436. http://dx.doi.org/10.3102/0034654310377086

Patrick, L., Care, E., & Ainley, M. (2011). The relationship between vocational interests, self-efficacy, and achievement in the prediction of educational pathways. *Journal of Career Assessment, 19,* 61–74. http://dx.doi.org/10.1177/1069072710382615

Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology, 106,* 315–329. http://dx.doi.org/10.1037/a0034027

Perez-Felkner, L., McDonald, S.-K., & Schneider, B. (2014). What happens to high-achieving females after high school? Gender and persistence on the postsecondary STEM pipeline. In I. Schoon & J. S. Eccles (Eds.), *Gender differences in aspirations and attainment* (pp. 285–320). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139128933.018

Porter, R. E., Fusarelli, L. D., & Fusarelli, B. C. (2015). Implementing the common core: How educators interpret curriculum reform. *Educational Policy, 29,* 111–139. http://dx.doi.org/10.1177/0895904814559248

R Development Core Team. (2014). R. [Software]. Vienna, Austria: R Foundation for Statistical Computing.

Reeve, C. L., & Heggestad, E. D. (2004). Differential relations between general cognitive ability and interest-vocation fit. *Journal of Occupa-*

tional and Organizational Psychology, 77, 385–402. http://dx.doi.org/10.1348/0963179041752673

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology, 107,* 645–662. http://dx.doi.org/10.1037/edu0000012

Rolfhus, E. L., & Ackerman, P. L. (1996). Self-report knowledge: At the crossroads of ability, interest, and personality. *Journal of Educational Psychology, 88,* 174–188. http://dx.doi.org/10.1037/0022-0663.88.1.174

Rounds, J., & Su, R. (2014). The nature and power of interests. *Current Directions in Psychological Science, 23,* 98–103. http://dx.doi.org/10.1177/0963721414522812

Schäfer, J. L. (1997). *Analysis of incomplete multivariate data: Monographs on statistics and applied probability* (Vol. 72). New York, NY: Chapman & Hall. http://dx.doi.org/10.1201/9781439821862

Scheerens, J., & Hendriks, M. (2014). State of the art of time effectiveness. In J. Scheerens (Ed.), *SpringerBriefs in Education: Effectiveness of time investments in education* (pp. 7–29). Cham, Germany: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-00924-7_2

Schiefele, U., Krapp, A., & Winteler, A. (1992). Interest as a predictor of academic achievement: A meta-analysis of research. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 183–212). Hillsdale, NJ: Erlbaum.

Schnabel, K. U., Alfeld, C., Eccles, J. S., Köller, O., & Baumert, J. (2002). Parental influence on students' educational choices in the United States and Germany: Different ramifications–Same effect? *Journal of Vocational Behavior, 60,* 178–198. http://dx.doi.org/10.1006/jvbe.2001.1863

Schoon, I., & Eccles, J. S. (Eds.). (2014). *Gender differences in aspiration and attainment: A life course perspective.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139128933

Schurtz, I. M., Pfost, M., Nagengast, B., & Artelt, C. (2014). Impact of social and dimensional comparisons on student's mathematical and English subject interest at the beginning of secondary school. *Learning and Instruction, 34,* 32–41. http://dx.doi.org/10.1016/j.learninstruc.2014.08.001

Schwanzer, A. D., Trautwein, U., Lüdtke, O., & Sydow, H. (2005). Entwicklung eines Instruments zur Erfassung des Selbstkonzepts junger Erwachsener [Development of a new instrument to assess self-concept of young adults]. *Diagnostica, 51,* 183–194. http://dx.doi.org/10.1026/0012-1924.51.4.183

Sells, L. W. (1980). Mathematics: The invisible filter. *Engineering Education, 70,* 340–341.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inferences.* Berkeley, CA: Houghton Mifflin.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Los Angeles, CA: Sage.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41,* 481–520. http://dx.doi.org/10.3102/1076998616646200

Strong, E. K. J. (1943). *Vocational interests of men and women.* Stanford, CA: Stanford University Press.

Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Frontiers in Psychology, 6,* 189. http://dx.doi.org/10.3389/fpsyg.2015.00189

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin, 135,* 859–884. http://dx.doi.org/10.1037/a0017364

Trautwein, U., Köller, O., Lüdtke, O., & Baumert, J. (2005). Student tracking and the powerful effects of opt-in courses on self-concept: Reflected-glory effects do exist after all. In H. W. Marsh, R. Craven, & D. M. McInerney (Eds.), *New frontiers for self research* (pp. 307–327). Greenwich, CT: Information Age Publishing.

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology, 98,* 788–806. http://dx.doi.org/10.1037/0022-0663.98.4.788

Trautwein, U., Neumann, M., Nagy, G., Lüdtke, O., & Maaz, K. (Eds.). (2010). *Schulleistungen von Abiturienten: Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand* [School achivement of upper secondary school students. The rearrangend upper secondary school on the trial]. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften. http://dx.doi.org/10.1007/978-3-531-92037-5

Updegraff, K. A., Eccles, J. S., Barber, B. L., & O'Brien, K. M. (1996). Course enrollment as self-regulatory behavior: Who takes optional high school math courses? *Learning and Individual Differences, 8,* 239–259. http://dx.doi.org/10.1016/S1041-6080(96)90016-3

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140,* 1174–1204. http://dx.doi.org/10.1037/a0036620

Wagner, W., Rose, N., Dicke, A.-L., Neumann, M., & Trautwein, U. (2014). Alle alles lehren—Schulleistungen in Englisch, Mathematik und den Naturwissenschaften vor und nach der Neuordnung der gymnasialen Oberstufe in Sachsen [Teaching everyone everything: Student achievement in English, mathematics and the natural sciences before and after Saxony's upper secondary school reform]. *Zeitschrift für Erziehungswissenschaft, 17,* 345–369. http://dx.doi.org/10.1007/s11618-014-0492-7

Watt, H. M. G., & Eccles, J. S. (Eds.). (2008). *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences.* Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/11706-000

Watt, H. M. G., Eccles, J. S., & Durik, A. M. (2006). The leaky mathematics pipeline for girls. *Equal Opportunities International, 25,* 642–659. http://dx.doi.org/10.1108/02610150610719119

Watt, H. M. G., Richardson, R. W., & Devos, C. (2013). (How) Does gender matter in the choice of a STEM teaching career and later teaching behaviours? *International Journal of Gender, Science and Technology, 5,* 187–206.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review, 30,* 1–35. http://dx.doi.org/10.1016/j.dr.2009.12.001

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25,* 68–81. http://dx.doi.org/10.1006/ceps.1999.1015

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology, 89,* 451–469. http://dx.doi.org/10.1037/0022-0663.89.3.451

Wigfield, A., Tonks, S., & Klauda, S. T. (2009). Expectancy-value theory. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 55–57). New York, NY: Routledge.

Young, T., & Lewis, W. D. (2015). Educational policy implementation revisited. *Educational Policy, 29,* 3–17. http://dx.doi.org/10.1177/0895904815568936

Received March 4, 2016
Revision received November 30, 2016
Accepted December 8, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 7, 1010–1028

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000187

Impacts of the CARE for Teachers Program on Teachers' Social and Emotional Competence and Classroom Interactions

Patricia A. Jennings
University of Virginia

Joshua L. Brown
Fordham University

Jennifer L. Frank, Sebrina Doyle, and Yoonkyung Oh
Pennsylvania State University

Regin Davis
Columbia University

Damira Rasheed and Anna DeWeese
Fordham University

Anthony A. DeMauro
University of Virginia

Heining Cham
Fordham University

Mark T. Greenberg
Pennsylvania State University

Understanding teachers' stress is of critical importance to address the challenges in today's educational climate. Growing numbers of teachers are reporting high levels of occupational stress, and high levels of teacher turnover are having a negative impact on education quality. Cultivating Awareness and Resilience in Education (CARE for Teachers) is a mindfulness-based professional development program designed to promote teachers' social and emotional competence and improve the quality of classroom interactions. The efficacy of the program was assessed using a cluster randomized trial design involving 36 urban elementary schools and 224 teachers. The CARE for Teachers program involved 30 hr of in-person training in addition to intersession phone coaching. At both pre- and postintervention, teachers completed self-report measures and assessments of their participating students. Teachers' classrooms were observed and coded using the Classroom Assessment Scoring System (CLASS). Analyses showed that CARE for Teachers had statistically significant direct positive effects on adaptive emotion regulation, mindfulness, psychological distress, and time urgency. CARE for Teachers also had a statistically significant positive effect on the emotional support domain of the CLASS. The present findings indicate that CARE for Teachers is an effective professional development both for promoting teachers' social and emotional competence and increasing the quality of their classroom interactions.

Keywords: teacher stress, mindfulness, teacher professional development, classroom interactions, emotion regulation

Supplemental materials: http://dx.doi.org/10.1037/edu0000187.supp

Understanding teachers' stress is critical for the stability and effectiveness of educational systems worldwide (Kyriacou, 2011). The most recent survey by MetLife (Markow, Macia, & Lee, 2013), with a demographically representative sample of 1,000 U.S. K–12 public school teachers, found that 59% of teachers reported being under great stress, a dramatic increase from 35% in 1985. There was also a

This article was published Online First February 13, 2017.

Patricia A. Jennings, Department of Curriculum, Instruction and Special Education, Curry School of Education, University of Virginia; Joshua L. Brown, Department of Psychology, Fordham University; Jennifer L. Frank, Department of Education, Pennsylvania State University; Sebrina Doyle, Bennett Pierce Prevention Research Center, Pennsylvania State University; Yoonkyung Oh, Department of Education, Pennsylvania State University; Regin Davis, Center for Public Research and Leadership, Columbia University; Damira Rasheed and Anna DeWeese, Department of Psychology, Fordham University; Anthony A. DeMauro, Department of Curriculum, Instruction and Special Education, Curry School of Education, University of Virginia; Heining Cham, Department of Psychology, Fordham University; Mark T. Greenberg, Department of Human Development and Family Studies, Pennsylvania State University.

The project described was supported by Award Numbers R305A120180 and R305A140692 from the Institute of Educational Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Institute of Educational Sciences or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Patricia A. Jennings, Curry School of Education, University of Virginia, P.O. Box 400273, 206D Bavaro Hall, Charlottesville, VA 22904. E-mail: tishjennings@virginia.edu

statistically significant decrease in teachers' self-reported job satisfaction from 62% in 2008 to 39% in 2012, the largest drop since 1984 when MetLife began the survey. These findings are consistent with a recent Gallup (2014) survey in which nearly half of K–12 teachers (46%) reported high daily stress during the school year, one of the highest stress levels among all occupational groups including nurses (46%) and physicians (45%).

Teacher stress and the resulting attrition are serious problems that negatively impact the quality of education, taking an emotional and psychological toll on school personnel and impacting student behavior and achievement (Greenberg, Brown, & Abenavoli, 2016; Hoglund, Klingle, & Hosan, 2015), particularly among high-poverty schools where both stress and attrition levels are the highest (Alliance for Excellent Education, 2014). Despite the documented high level of teacher stress, little research has addressed ways to reduce it. Developing and testing new approaches designed to help teachers manage the stresses of teaching and improve the quality of classroom interactions that promote student learning is critical to effectively supporting and maintaining the teaching workforce. Responding to this need, the current study examined the efficacy of the Cultivating Awareness and Resilience in Education (CARE for Teachers) professional development program.

Understanding Teacher Stress in the Classroom Context

There are numerous factors related to high levels of teacher stress and consequent burnout worldwide. These include managing student misbehavior, providing support to needy and/or unmotivated students, feeling that their workload is overwhelming, feeling a lack of control over decisions that affect them and their students, having little time to relax due to the need to take a great deal of work home, and feeling the constant pressure to be accountable for student outcomes (Richards, 2012). Indeed, levels of stress among teachers have increased in the current era of high stakes testing (Dworkin & Tobe, 2014). These factors can provoke strong negative emotions and teachers consistently report that coping with these emotions is a major stressor (Carson, Weiss, & Templin, 2010). Negative emotions may impair teachers' cognitive functioning and well-being, which can have a negative effect on instruction (Emmer & Stough, 2001). Frequently experiencing negative emotions may reduce teachers' intrinsic motivation and self-efficacy (Sutton & Wheatley, 2003). Long-term, constant emotional distress can impair teachers' performance leading to burnout (Tsouloupas, Carson, Matthews, Grawitch, & Barber, 2010), and increased student misbehavior (Osher et al., 2007). In contrast, teachers who manage their stress and effectively regulate their emotions can more frequently experience positive emotions, leading to greater resilience and enjoyment of teaching (Gu & Day, 2007).

Teachers who experience high levels of stress and frustration may transmit these feelings and their impacts directly to students via "stress-contagion" (Wethington, 2000, p. 234). Examining data from a nationally representative sample of first graders (N = 10,700), Milkie and Warner (2011) found that children in classrooms in which teachers reported experiencing greater levels of stress had higher internalizing and externalizing disorders. Similarly, in a Canadian sample of 406 elementary school students and their teachers (N = 17), Oberle and Schonert-Reichl (2016) found that teachers' self-reported burnout was linked to students' physiological stress regulation as measured by the diurnal pattern of cortisol.[1] Higher levels of teacher burnout significantly predicted the variability in students' morning cortisol levels suggesting evidence of an impaired stress response.

A meta-analysis of 65 independent studies of teacher stress drawn from international sources of literature identified improved emotion regulation as a key to preventing teacher stress (Montgomery & Rupp, 2005). The emotional labor teachers expend managing negative emotions may result in emotional exhaustion, a risk factor for burnout (Chang, 2009) and developing adaptive coping strategies may support teachers' well-being and performance (Chang, 2013).

Jennings and Greenberg (2009) presented the prosocial classroom theoretical model and proposed that certain social and emotional competences support teachers' ability to cope with the demands of teaching and prevent burnout. These competencies include self-awareness of emotional states and cognitions and the ability to effectively regulate their emotions while teaching to avoid becoming emotionally depleted and maintain their emotional energy to effectively respond to students' needs. According to the model, when teachers lack the social and emotional competences required to manage the demands of teaching, their well-being erodes and leads to a deterioration of the classroom climate and teacher stress, triggering a "burnout cascade" (p. 492). In contrast, teachers with high levels of social and emotional competences are able to cope with the demands of the classroom, maintain a positive classroom climate, build and maintain supportive relationships with their students, and establish consistent classroom interactions that promote student learning.

Empirical research has begun to show support for this model. For example, a randomized controlled study of the Head Start REDI model found that preschool teachers who received training and weekly mentoring support showed improvements in emotional supportiveness, as measured by the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008), compared with control teachers (Domitrovich et al., 2009). The CLASS is a well-validated and commonly used observational measure of classroom interaction quality that assesses emotional support, classroom organization and instructional support. The REDI training included instruction on a social and emotional learning curriculum and emphasized the importance of generalization of social and emotional learning through extension activities and teaching and classroom management strategies. Similarly, a cluster randomized controlled trial of the RULER social and emotional learning program delivered in fifth- and sixth-grade classrooms targeting both teachers' and students' emotional knowledge, self-awareness, and self-regulation skills, found statistically significant program impacts after two years on classroom interaction quality as measured by the emotional support, instructional support, and classroom organization domains of the CLASS (Hagelskamp, Brackett, Rivers, & Salovey, 2013).

[1] The typical diurnal cortisol cycle involves a burst of secretory activity following awakening with a diurnal decline across the day. A disrupted diurnal cortisol cycle may be evidence of an impaired stress response (Collomp et al., 2016).

Mindfulness-Based Interventions

One method for reducing stress and promoting emotional awareness and self-regulation is through engaging in mindful awareness practices. Various mindful awareness practices have been combined to create mindfulness-based interventions (MBIs). MBIs were popularized as an approach to stress reduction through the work of Jon Kabat-Zinn's (1982) mindfulness-based stress reduction (MBSR) program. Kabat-Zinn (2003) defined *mindfulness* as "the awareness that emerges through paying attention on purpose, in the present moment, and nonjudgmentally to the unfolding of experience moment by moment" (p. 144). In order to further refine this definition for research purposes, Bishop et al. (2004) conceptualized mindfulness as involving two primary dimensions: (a) directing one's attention to the present moment and (b) cultivating an orientation to one's experience marked by curiosity, openness, and acceptance.

Although adapted from practices from a variety of religious traditions, secular mindful awareness practices do not involve religious belief, language, or ritual and the rationale for engaging in such practices is grounded in research. This is particularly important for applications designed for use in public school settings (Jennings, 2016b). Mindfulness can be cultivated through a variety of practices including mindfulness meditation, yoga, tai chi, and Qigong practices and can be practiced formally or informally, such as during routine daily activities like walking, eating, and listening in a mindful state (Williams & Kabat-Zinn, 2011). Over the past decade there has been a rapid growth of mindfulness-based programming delivered in public school settings for both teachers and students (Felver & Jennings, 2016). However, little rigorous research has evaluated its efficacy to reduce teacher stress and improve the quality of interactions between teachers and students in classrooms.

Empirical reviews of MBIs have shown psychological and physiological improvements in clinical and nonclinical adult populations such as reduced stress, anxiety, and depression and increased well-being (Eberth & Sedlmeier, 2012; Khoury et al., 2013; Sharma & Rush, 2014). Considerable research has examined the underlying neurophysiological effects of mindful awareness practices, specifically as they relate to emotion regulation (Corcoran, Farb, Anderson, & Segal, 2010). For example, Hölzel and colleagues (Hölzel et al., 2011, 2013; Tang, Hölzel, & Posner, 2015) found that participants in an 8-week MBSR program showed increased gray matter and brain density in the hippocampus, an area of the brain associated with emotion regulation, compared with wait-list controls.

Effects of Mindfulness-Based Interventions on Teacher Stress

Mindful awareness practices may be particularly useful for helping teachers develop the skills they need to manage the demands of teaching. These practices may promote adaptive emotion regulation and coping which may lead to declines in stress, burnout and distress, and more energy and self-regulatory resources (e.g., more joy, more satisfaction, more well-being) that can then be invested in supportive teacher-student interactions that promote student learning (Roeser, 2016; Roeser, Skinner, Beers, & Jennings, 2012; Skinner & Beers, 2016).

Recently, randomized controlled studies have begun to investigate causal relationships between MBIs and stress reduction among teachers (Crain, Schonert-Reichl, & Roeser, 2016; Kemeny et al., 2012; Roeser et al., 2013) and improvements in classroom interactions (Flook, Goldberg, Pinger, Bonus, & Davidson, 2013). The current study examines the effects CARE for Teachers, which introduces emotion skills instruction, mindful awareness and stress reduction practices and caring and listening practices to promote improved emotion regulation, teaching efficacy and mindfulness and to reduce psychological and physical distress.

The first two pilot studies of CARE for Teachers examined program feasibility and attractiveness and preliminary evidence of efficacy (Jennings, Snowberg, Coccia, & Greenberg, 2011). The first study involved 31 educators from a high-poverty urban setting. The second study involved student teachers and 10 of their mentor teachers working in suburban/semirural schools ($N = 43$). Although educators working in the urban schools showed significant pre–post improvements in mindfulness and time urgency, the suburban/semirural sample did not, suggesting that CARE may be more efficacious in supporting teachers working in high-risk settings.

In a pilot study of the initial efficacy of CARE for Teachers, teachers were randomly assigned to CARE for Teachers ($n = 23$) or a wait-list control group ($n = 27$) and assessed pre- and postintervention on self-report measures to assess their emotion regulation, burnout, mindfulness, and teaching efficacy (Jennings, Frank, Snowberg, Coccia, & Greenberg, 2013). Compared with controls, teachers who received CARE for Teachers demonstrated statistically significant improvements in emotion regulation, mindfulness, and teaching efficacy, and reductions in time-related stress and physical symptoms associated with stress.

Two studies have examined another MBI model designed for teachers, the Stress Management and Relaxation Techniques in Education (SMART) program. The first study randomly assigned teachers ($n = 38$) and parents ($n = 32$) of students with disabilities to receive the SMART intervention or waitlist control group (Benn, Akiva, Arel, & Roeser, 2012). Compared with the control group, SMART participants showed decreased stress and anxiety and increased mindfulness, self-compassion, personal growth, empathy, and forgiveness. Results also showed participants' mindfulness at postintervention mediated treatment effects on stress, anxiety, negative affect, and personal growth measured at a 2-month follow-up.

The second SMART trial involved two samples of elementary and secondary public school teachers, one in the U.S. and one in Canada (Roeser et al., 2013). One hundred and 13 teachers were randomly assigned to SMART or to a wait-list control group and were assessed at pretest, postintervention, and at a 3-month follow-up using self-report measures and physiological indicators of stress including salivary cortisol (Canada only), blood pressure, and resting heart rate. The Canadian sample was also assessed on attentional abilities and working memory using a computer task-based assessment. At posttest, teachers receiving SMART showed decreased occupational stress and burnout, as well as increased mindfulness and self-compassion, compared with control group teachers. In the Canadian sample, teachers receiving SMART also showed improvements in attentional abilities and working memory. No statistically significant intervention effects were found on physiological indicators of stress. Results at the 3-month follow-up

indicated changes in mindfulness and self-compassion at posttest mediated SMART participants' stress, burnout, anxiety, and depression at follow-up.

Another study involving the same sample found that teachers randomized to SMART reported improved mood at work and home and improvements in the amount and quality of sleep (Crain et al., 2016). Intervention-related group differences in mindfulness and rumination (excessive worry) at postintervention partially mediated the reductions in negative mood and increases in sleep quality at 3-month follow-up.

A small pilot study (N = 18) examined the effects of MBSR adapted for teachers on psychological distress, mindfulness, self-compassion, burnout, neuropsychological and attentional task performance, diurnal cortisol and observations of interaction quality (Flook et al., 2013). Pre–post comparisons showed that intervention teachers (n = 10) showed statistically significant reductions in psychological symptoms and burnout and increases in self-compassion. They also showed improvements in performance on a computer task of affective attentional bias and observer-rated classroom organization. In contrast, the teachers assigned to the control condition (n = 8) showed statistically significant declines in diurnal cortisol functioning.

The results of MBIs specifically designed for teachers show promise for reducing teachers' occupational stress, promoting social and emotional competencies, and improving the quality of their classroom interactions. However, interpretation and generalizability have been limited by small samples (Beshai, McAlpine, Weare, & Kuyken, 2016; Flook et al., 2013; Franco, Manas, Cangas, Moreno, & Gallego, 2010; Frank, Reibel, Broderick, Cantrell, & Metz, 2015; Jennings et al., 2013; Poulin, Mackenzie, Soloway, & Karayolas, 2008; Taylor et al., 2016a, 2016b) and no studies to date have accounted for potential school context effects by employing analytic methods appropriate to the multilevel structure of such data in which teachers/classrooms are clustered within schools. Thus, the current study builds on and advances this research by (a) including the largest sample of teachers in an MBI impact study to date, and one that is drawn from a large inner city school district in the U.S., with substantial racial/ethnic diversity, and (b) randomizing teachers within schools and using analytic methods that account for the clustering of teachers and classrooms within schools.

The CARE for Teachers Logic Model

CARE for Teachers is specifically designed to address teachers' social and emotional competences as hypothesized in the CARE for Teachers logic model (see Figure 1). In the present study, the population of focus was K–5 teachers. The CARE for Teachers program elements of emotion skills instruction, mindful awareness and stress reduction practices and caring and listening practices are hypothesized to promote increases in adaptive emotion regulation, teaching efficacy and mindfulness and reductions in psychological and physical distress as well as improvements in classroom interactions that promote learning (e.g., emotional support and classroom organization). The program elements are hypothesized to have a synergistic effect on the hypothesized outcomes such that no one single program element is hypothesized to have a unique and direct impact on any one outcome. A similar logic model was developed and tested in previous studies of CARE for Teachers (Jennings et al., 2011, 2013). For the current study we refined the model slightly in response to previous work.

We hypothesized that teachers randomly assigned to receive CARE for Teachers would show statistically significant improvements in adaptive emotion regulation, teaching efficacy, and mindfulness and reductions in psychological distress and physical distress, compared with teachers randomly assigned to the waitlist condition. We also hypothesized that teachers trained in CARE for Teachers would promote classroom interactions that exhibit higher

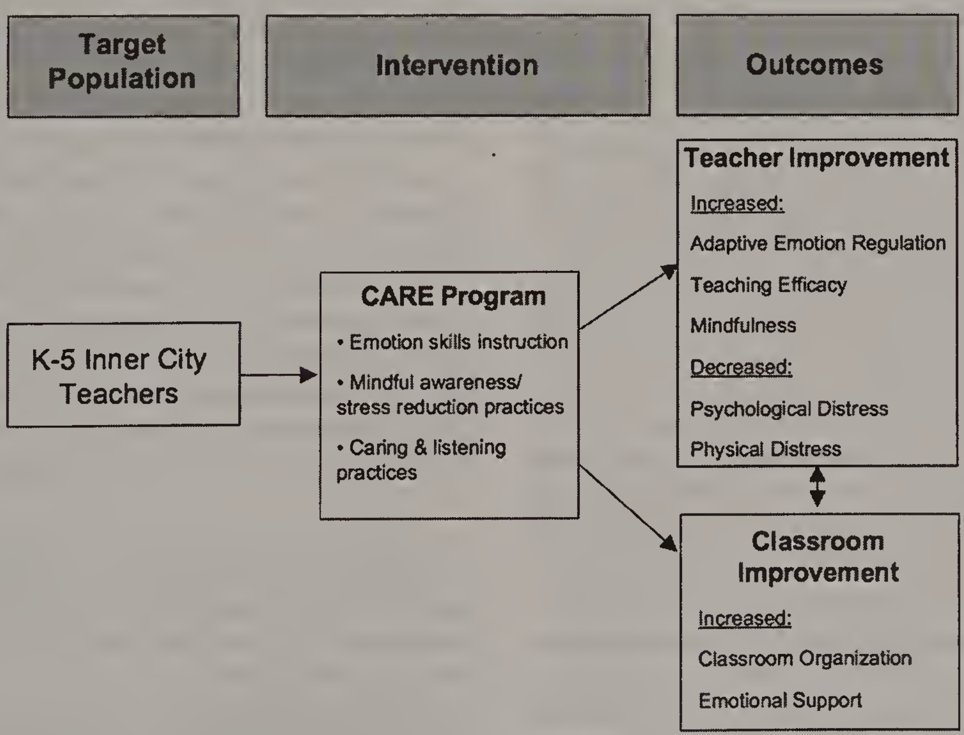


Figure 1. CARE for Teachers logic model.

levels of emotional support and classroom organization than the classrooms of teachers randomly assigned to the waitlist control group. Finally, we hypothesized that teachers who participated in CARE for Teachers would perceive the program as having high social importance and acceptability (e.g., high social validity).

Method

Procedures

Recruitment. School recruitment took place in Spring of 2012 (Cohort 1; C1) and 2013 (Cohort 2; C2) among inner city public K–5 elementary schools in a high poverty region of New York City (the Bronx and Upper Manhattan). We chose elementary schools because teachers at this level spend most of the day with the same group of students and are thus able to have greater influence on observable dimensions of classroom interactions than teachers in secondary schools. High poverty schools were chosen for this study because the results of previous research indicated that CARE for Teachers was most helpful for teachers working in these contexts (Jennings et al., 2011).

Schools were initially recruited by approaching principals of schools in the targeted regions, explaining the CARE for Teachers program and the purpose of the research, and inviting them to participate. In a written memorandum of understanding, principals in each participating school agreed to: support enrollment and participation of at least four teachers per school, help facilitate scheduling of research activities, and support distribution of study information to parents. Principals also agreed to release participating teachers to participate in the CARE for Teachers program during paid work time and to cover the cost of a substitute for each participating teacher for one training day. The study enrolled 36 of the 73 schools approached. Factors that inhibited school recruitment were largely due to principals already having too many programs, too few eligible (see below) teachers, or lack of interest.

Within the 36 participating schools, eligible teachers were identified that met the following criteria: taught in a classroom within the K–5 range of grade levels, taught general education (e.g., no art or physical education teachers), lead teacher in his or her classroom (e.g., no cotaught classrooms),[2] taught the same students for the entirety of the school day, and had classrooms that were representative of the average classroom in this city (e.g., no single gender classrooms). All eligible teachers were invited by their principals to attend recruitment meetings for CARE for Teachers led by the study principal investigators with the support of research staff. The program was described during recruitment as follows:

> CARE is an innovative professional development program that introduces specific skills to help teachers manage stress and improve their teaching effectiveness. CARE combines emotion skills training with mindfulness-based stress reduction activities and provides teachers with opportunities to practice applying these skills in the classroom.

During the meeting, study and teacher participation requirements were described in detail; contact information was collected for all attendees. The following fall, research staff contacted all teachers who attended initial recruitment meetings to complete the consent process. As a result of these efforts, 1,084 teachers were assessed for eligibility, 491 of these did not meet the study inclusion criteria and 68 could not participate for other reasons, leaving 525 eligible teachers.[3]

Sample. Of the 525 eligible teachers approached for participation, 301 declined to participate resulting in a sample of 224 teachers recruited from 36 schools, a 43% response rate (*Mdn* = 6, range = 2–10 teachers). C1 consisted of 53 teachers from 8 schools, and C2 consisted of 171 teachers from an additional 28 schools. Attrition was low at 6% (five from control, eight from intervention) at posttest. Ninety-three percent of participants were female ($n = 209$) and 7% were male ($n = 15$). The sample was racially and ethnically diverse with 74 teachers (33%) identifying as White, 69 (31%) as Hispanic, 59 (26%) as African American/Black, 10 (5%) as Asian, and 12 (5%) identifying as being of a mixed racial background. Teachers' ages ranged from 22–73 years (*Mdn* = 40) and number of years teaching ranged from 0 to 32 years. Ninety-six percent had a Master's/Specialist degree ($n = 213$) or Doctoral Degree ($n = 1$). Active consent was obtained from teachers in accordance with both the University's and district's institutional review board procedures.

Compared with the statistics available for New York City Independent Budget Office in 2014 the sample had more females compared with the general population of elementary/middle school teachers (NYC = 84%). There were fewer White teachers (NYC = 59%) and more Hispanic (NYC = 19%) and African American/Black (NYC = 20%) teachers than the general population of New York City teachers. Participating teachers were similar in age ($M = 41.5$ vs. $M = 40$); however, they reported more years of teaching experience ($M = 12.5$ vs. $M = 10.6$ years). No statistics were available to determine how closely the study sample of teachers matched the general population of New York City teachers with regard to percent holding graduate degrees.

Participants were distributed across grades with 39 (17%) teaching Kindergarten, 40 (18%) in 1st grade, 33 (15%) in 2nd grade, 36 (16%) in 3rd grade, 33 (15%) in 4th grade, 40 (18%) in 5th grade, and three (1%) in multiple grades (one K–1, one 2–3 and one 3–4 combo). One-hundred and 90 (85%) participants were general education teachers, 30 (13%) were teaching in combined language (bilingual, ESL, ELL or dual) classes, and four (2%) teachers endorsed teaching in a special education inclusion classroom as a general education teacher teaching alone (e.g., not coteaching with a special education teacher). Class sizes (the average number of students in the classroom across two observation days) were slightly below the 2014 district average ($M = 23.67$ vs. $M = 25.19$) although there was considerable variation (range = 13–33).

Randomization. The present study evaluated the efficacy of the CARE for Teachers intervention for K–5 teachers and classrooms using a two-level (teachers/classrooms, schools) multisite cluster randomized trial design with intervention at level two (teachers) and schools serving as naturally occurring blocks. Ran-

[2] At the time of recruitment, the New York City schools were beginning to transition to a new model to support efforts to include students with disabilities in the general education classrooms involving special education teachers co-teaching with general education teachers. Due to the limitations of our research design, we could only recruit teachers working in classrooms without a co-teacher.

[3] A CONSORT flow diagram representing the progress through the phases of the present randomized controlled trial and a table reporting on participant attrition are provided in the online supplemental materials.

domization of teachers to CARE for Teachers or the waitlist control group was conducted after baseline data collection by schools and by grade for each cohort. We utilized a block randomization method to randonize participants into groups of approximate equal sample size within schools. This was achieved by establishing a set block size for each school, and then generating all possible balanced combinations of assignments within the block using a computer generated random-number sequence, with a new random-number seed introduced for each iteration. Randomized blocks were then randomly chosen to determine participants' assignment to groups resulting in 118 teachers assigned to receive the CARE for Teachers program and 106 assigned to the wait-list control condition.

We randomized teachers within schools to ensure that teacher assignment was balanced across grade level. Compared with school-randomized designs, randomly assigning teachers within schools has been recommended as a strategy to control between-school variability (Werthamer-Larsson, 1994), and requires fewer schools to achieve adequate statistical power to detect small to moderate effects (Blitstein, Hannan, Murray, & Shadish, 2005; Cornfield, 1978; Raudenbush, Martinez, & Spybrook, 2007; Schochet, 2008). The CARE for Teachers intervention is entirely teacher-focused and does not presume synergistic influences afforded by whole school implementation. While the chosen design posed a potential threat to the internal validity of the experiment due to possible contamination or spillover of program effects from intervention to control group teachers within a school, we decided that the within school randomization was still the preferable design choice and that contamination would be highly unlikely. Furthermore, sharing their experiences of CARE for Teachers program activities with nontrained colleagues would not be likely to provide the necessary detail, scaffolding of learning, and intensity afforded by the direct experience of group participation in the sequenced program activities to effect statistically significant changes in control group teachers. Therefore, we anticipated the risk of contamination would be well below the approximately 50% threshold at or beyond which the random assignment of schools instead of teachers within schools would be preferable (Rhoads, 2011).

Teachers assigned to the intervention condition received CARE for Teachers in the Fall/Winter of 2012–2013 for C1 and 2013–2014 for C2 immediately following initial data collection and randomization. These teachers also received standard professional development activities as assigned by their schools with the exception of one CARE for Teachers training day: on this day intervention teachers received the CARE for Teachers program rather than the standard professional development delivered to all other teachers, including control teachers. Teachers in the wait-list control condition only received standard professional development activities as assigned by their schools. With the exception of time spent in professional development related to stress reduction, mindfulness, or other meditative activities (e.g., CARE for Teachers), no statistically significant differences were found between groups on amount of professional development (i.e., curriculum/academic instruction, student/classroom behavior, and social and emotional learning) received during the intervention school year. Teachers in the control condition were offered CARE for Teachers following the completion of all research activities for their cohort. Of the control teachers, 51% completed the CARE training (n = 54).

Intervention/CARE for Teachers. The CARE for Teachers program model is a comprehensive system designed to reduce teachers' stress and to promote and support teachers' social and emotional competences over the course of one full school year. Following best practices in adult learning, CARE for Teachers introduces material sequentially, utilizing a blend of didactic, experiential, and interactive learning processes. The program presents a structured set of mindful awareness practices including breath awareness practice, mindful walking and stretching, listening and compassion practices, as well as didactic and experiential practices to promote emotion awareness and emotion regulation (see Jennings et al., 2011, 2013, and Jennings, 2016a, for more extensive descriptions of the CARE for Teachers program model).

CARE for Teachers was delivered in 30 hr over 5 in-person training days (6 hr each) between November and February; the first two training days were offered back-to-back in November (one of these days was a designated professional development day for all teachers), and then two training days were offered in the subsequent month separated by several weeks. The breaks in between sessions gave teachers an opportunity for practice, reflection, and application of the material to their teaching. Each CARE for Teachers training was presented by a team of three facilitators who met a standard set of requirements, including a minimum of a master's degree in education, psychology or related area, a minimum of two years' experience with the program, and a personal mindfulness practice.

Over 90% of the participants attended at least 4 of the 5 days (M = 4.49) of the program. All participants received a program workbook, along with an audio CD/MP3 of recorded mindful awareness practices to facilitate home practice. In addition to in-person sessions, teachers were scheduled to receive a series of three one-on-one phone coaching calls (DeWeese et al., in press). Each participating teacher was assigned to a specific coach for the duration of the program. Coaches were either facilitators or training fidelity coders who had completed at least one CARE for Teachers training. Coaching calls were offered during intersession breaks following Days 2, 3, and 4; on average the calls lasted 26 min (range = 9–60 min) and were intended to support teachers' development of personal mindful awareness practices and the application of CARE for Teachers skills and concepts to their teaching. Participants completed a CARE for Teachers practices questionnaire either before or during the coaching call. Coaches discussed with participants their use of practices, what they found helpful, and whether they had any questions or challenges for which they needed support. Coaching calls were conducted regardless of participants' attendance at sessions; a brief review of material was provided if a participant was absent for the session prior to a given call.

Teachers were compensated at the district approved training rate of $19.12 an hour for one 6-hr training day that occurred on the weekend. Schools were compensated for substitute teacher pay for two training days scheduled while school was in session. Schools covered the cost for one day of substitute teacher pay. No compensation was provided to schools or teachers for the one training day offered during the regularly scheduled in-service professional development day.

Fidelity and quality. Two aspects of implementation were assessed: fidelity and quality. Fidelity was assessed by two trained fidelity coders for all CARE for Teachers sessions using the CARE Daily Session Rating Forms (Doyle, Jennings, DeWeese, & Frank, 2014). The Daily Session Rating Form is an observational measure that assessed the completion of program components and how well the participant learning objectives were met. Codes were checked for reliability and disagreements were rectified by consensus with support from the coding supervisor. On average, 88% (range = 86–91%) of the facilitation activity components listed in the manual were completed. Interrater reliability for component measurement was acceptable (κ = .67; Cohen, 1960). Completion of participant learning objectives for each activity was rated on a 0–4 scale. Participant objectives were met at an adequate to exemplary level (M = 3.43, range = 3.29–3.65). Interclass correlation ratings for "objectives met" were excellent (.75).

The quality of facilitation skill was coded using the CARE Facilitator Rating Form, a modified version of the Iowa Strengthening Families Program Facilitator Delivery Ratings (Iowa State University Extension and Outreach, 2010). Coders provided ratings each day on 10 positive (e.g., engaging participants, explaining material well) and six negative (e.g., losing track of time, being critical of participants) facilitation skills (rated on a 0–4 scale). Overall, facilitators demonstrated a high level of positive and low level of negative facilitation skills (M = 3.77). Interclass correlation ratings for facilitation skill were excellent (.79).

Data Collection

Self-report and report on student assessments. Participants completed an online battery of self-report measures and assessments of the students in their class prior to the intervention in fall and again in spring of the same school year. Measure items were grouped by measure and were not randomized. Teachers were compensated for survey completion during afterschool hours equivalent to the district pay rate of $42 an hour. The questionnaires at each time point took approximately 45 min to complete.

Classroom observations. Observations of the overall quality of interactions between teachers and students were conducted by trained, independent observers in the classroom in both the fall (preintervention) and spring of the school year using CLASS (Pianta et al., 2008). The K–3 version of the CLASS was used for all classrooms (K–5) to maintain measurement consistency across all classrooms. Observations were conducted by 24 ethnically diverse certified coders who were blind to teacher intervention condition. In addition to required certification in the CLASS, observers also received live training, and participated in regular calibration meetings and midpoint reliability checks. Two observations of each participating teachers' classroom were conducted at both pre- and postassessment. Observations took place on two separate days within the same week for approximately one hour each day while the target teacher was instructing the class. Each observation day consisted of three 22-min cycles; each cycle was comprised of a 15-min interval of observing CLASS indicators and a 7-min coding period. Observers were randomly assigned to each observation day; different observers coded the first and second day at pre- and postassessment to control for coding bias due to prior exposure. Thirty-three percent of the 867 total observations were double-coded across pre- and posttest. No compensation was provided to teachers for classroom observations.

Measures

Measures were selected based on our previous research (and other research on MBIs with teachers and other adult populations) and the CARE for Teachers logic model proposing that the program has direct effects on teachers' adaptive emotion regulation, teaching efficacy, mindfulness, psychological distress, physical distress and the quality of classroom emotional support and organization.

Self-report and assessment of students. Participants completed self-report measures to assess adaptive emotion regulation, teaching efficacy, mindfulness, psychological distress and physical distress. Teachers assessments of their students were collected at the same time via the same online system (e.g., proportion of students with IEPs or 504 plan, ever suspended, and average learning support at home). Coefficient alphas for self-report scales were computed for all measures at pre- and posttest. Ranges of coefficient alphas at both time points are provided for each measure below.

Adaptive emotion regulation. One measure, the Emotion Regulation Questionnaire (ERQ; Gross & John, 2003), was used to assess teachers' adaptive emotion regulation. This 10-item scale assesses two emotion regulation strategies: cognitive reappraisal and expressive suppression. Respondents reported on emotional experience ("what you feel like inside") and emotional expression ("how you show your emotions in the way you talk, gesture, or behave") on a 7-point Likert-type scale (1 = *strongly disagree* to 7 = *strongly agree*). Coefficient alphas ranged from .67 to .68.

Teaching efficacy. One measure, the Teachers' Sense of Efficacy Questionnaire-Short Form, was used to assess teaching efficacy (TSES; Tschannen-Moran & Woolfolk Hoy, 2001). This short form is a 12-item measure of three dimensions of teaching efficacy: *efficacy for instructional strategies* (e.g., "How much can you use a variety of assessment strategies?"), *efficacy for classroom management* (e.g., "How well can you keep a few problem students form ruining an entire lesson?"), and *efficacy for student engagement* (e.g., "How much can you do to foster student creativity?"). Items asked teachers to indicate "how much they can do" in response to various classroom and instructional challenges; items were rated on a 9-point Likert scale (1 = *nothing* to 9 = *a great deal*). Coefficient alphas ranged as follows: efficacy for instructional strategies = 0.85, efficacy for classroom management = .83–.85, and efficacy for student engagement = .78–.83.

Mindfulness. Two measures assessed general mindfulness and mindfulness as it applies to classroom interactions. The first measure used was The Five Facet Mindfulness Questionnaire (FFMQ; Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006). This 39-item instrument has five subscales: *observing* (e.g., "I pay attention to how my emotions affect my thoughts and behavior"), *describing* (e.g., "Even when I'm feeling terribly upset, I can find a way to put it into words"), *acting with awareness* (e.g., reverse item: "I find myself doing things without paying attention"), *nonjudgmental* (e.g., reverse item: "I tell myself I shouldn't be feeling the way I'm feeling"), and *nonreactive* (e.g., "When I have distressing thoughts or images, I feel calm soon after"). Respondents were asked to indicate the extent to which various

mindfulness-related statements are generally true for them; items were rated on a 5-point Likert scale (1 = *never or rarely true* to 5 = *very often or always true*). Coefficient alphas for the subscales ranged as follows: *observing* = .83–.85; *describing* = .89–.91; *acting with awareness* = .89–.91; *nonjudgmental* = .85–.92; and *nonreactive* = .74–.77.

The second measure used was the 5-item interpersonal mindfulness subscale of the Mindfulness in Teaching Scale (MTS; Frank, Jennings, & Greenberg, 2016). Items are focused on mindfulness as it applies to classroom interactions (e.g., reverse item: "I am often so busy thinking about other things that I am not really listening to my students"). Items are answered on a 5-point Likert-type scale indicating how true each statement is for the respondent (1 = *never true* to 5 = *always true*). The coefficient alphas for *interpersonal mindfulness* ranged from .66 to .72.

Psychological distress. Seven measures were used to assess teachers' psychological distress. The first measure used was the Patient Health Questionnaire 8-item Depression Scale (PHQ-8; Kroenke et al., 2009). This 8-item measure of depressive symptoms (e.g., "feeling down, depressed, or hopeless") is rated on a 4-point Likert-type scale (1 = *not at all* to 4 = *nearly every day*). The coefficient alphas for the PHQ-8 was .87 at pre- and posttest.

The second measure used was the Generalized Anxiety Disorder 7-item Scale (GAD-7; Spitzer, Kroenke, Williams, & Löwe, 2006) from the Patient Health Questionnaire. It measures generalized anxiety symptoms (e.g., "feeling nervous, anxious, or on edge") on a 4-point Likert-type scale (1 = *not at all* to 4 = *nearly every day*). The coefficient alphas for the GAD-7 ranged from .92 to .93.

The third measure used for psychological distress was the International Positive and Negative Affect Rating Short Form (PANAS; Thompson, 2007). This brief 10-item measure asks participants to rate how they "felt during the past few weeks" on 10 emotions using a 5-point Likert-type scale (1 = *very little or not at all* to 5 = *extremely*). Coefficient alphas for the positive and negative affect subscales ranged from .75 to .92.

The forth measure used was the Patient Reported Outcomes Measurement Information System Sleep Disturbance Questionnaire (PROMIS; Buysse et al., 2010). This 4-item scale asks participants to rate the quality of their sleep and sleep patterns over the past 7 days (e.g., "My sleep quality was refreshing") on a 5-point Likert-type scale (1 = *not at all* to 5 = *very much*). The coefficient alphas for the PROMIS ranged from .85 to .87.

The fifth measure is the Emotional Exhaustion subscale of the Maslach Burnout Inventory–Educators' Survey (MBOI; Maslach, Jackson, & Leiter, 1997). This subscale measures burnout syndrome in teachers, (e.g., "I feel emotionally drained from my work") on a 7-point Likert-type scale (1 = *never* to 7 = *every day*). Coefficient alphas for the *emotional exhaustion* subscale were .91 at pre- and posttest.

The sixth measure of psychological distress used is the Perceived Stress Scale (PSS; Cohen, Kamarck, & Mermelstein, 1983). The PSS is a 4-item scale that assesses how difficult stressors were to handle over the last month (e.g., "How often have you felt that you were unable to control the important things in your life?"). Items are rated on a 5-point Likert-type scale (1 = *never* to 5 = *very often*). The coefficient alphas for the PSS ranged from .77 to .78.

The final scale used is the Time Urgency Scale (TUS; Landy, Rastegary, Thayer, & Colvin, 1991). The TUS assesses the multidimensional construct of time pressure (e.g., time-related stress). The subscales measure *speech patterns* (five items such as "I talk more rapidly than most people"), *eating behavior* (five items such as "I eat rapidly, even when there is plenty of time"), *competitiveness* (six items such as "I go all out'"), *task-related hurry* (three items such as "I usually work fast"), and *general hurry* (five items such as "I often feel very pressed for time"). Respondents were asked to indicate the extent to which various descriptors applied to them personally on a 5-point Likert-type scale (1 = *strongly disagree* to 5 = *strongly agree*). Coefficient alphas for the subscales ranged as follows: *speech patterns*, .70–.75; *eating behavior*, .85; *competitiveness*, .73–.74; *general hurry*, .73–.82; and *task-related hurry*, .54–.65.

Physical distress. Two measures were used to assess teachers' physical distress. The first measure is the Gastrointestinal and General Aches subscales Daily Physical Symptom Checklist (DPS; Larsen & Kasimatis, 1991). Participants were asked to indicate (yes/no) whether they experienced each particular symptom "today." Symptoms included pain such as headache and backache and gastrointestinal problems such as nausea and diarrhea. A sum score was created for each subscale; coefficient alphas ranged as follows: gastrointestinal = .55–.58, aches = .56–.63.

The second measure of physical distress focused on participant medication use. Participants were asked to indicate (yes/no) whether they were currently taking medications for 12 different common conditions (e.g., hypertension, heart condition, hormone replacement). A sum score for medication use was created. Coefficient alphas ranged across pre- and posttest from .27–.28; low alphas are expected as most conditions were not expected to correlate.

Teacher reports on student assessments. To assess the proportion of students in each class with an IEP or 504 plan, teachers were asked to respond (yes/no) to the following question, "Does this child have an IEP or 504 plan?" To assess the proportion of students in each class who had ever been suspended, teachers were asked to respond (yes/no) to the following question, "Has this child ever been suspended from school because of misbehavior?" If data was found missing on either of these items, we substituted the missing data with data from school records. To assess the average level of home support for learning of students in each class, teachers were asked to respond to the following question about each of their students using a 4-point Likert-type scale, "How would you characterize the level of support for learning in this child's home?" Responses ranged from 1 = *very poor* to 4 = *very good.*

Classroom observations. The CLASS (Pianta et al., 2008) assesses interactions between teachers and students and can be grouped into three domains of quality: *emotional support* (comprised of four dimensions: positive climate, negative climate, teacher sensitivity, and regard for student perspective), *classroom organization* (comprised of three dimensions: behavior management, productivity, and instructional learning formats), and *instructional support* (comprised of three dimensions: concept development, quality of feedback, and language modeling). Although we hypothesized that CARE for Teachers would impact the domains of *emotional support* and *classroom organization*, we included all three domains in our coding protocol to maintain measure validity as previous research on the validity and reliability of the CLASS included all three domains in the coding protocol.

Observers rated the CLASS dimensions (1 = *very low* to 7 = *very high*) during the three observational cycles on two days at each data collection wave. Scores were averaged across the coding cycles within dimension and then within domain. Thus, a participant's observation score is based on the average of all 15-min observations collected at pre- and posttest; observers' scores from double-coded classrooms were averaged to create one score.

The internal reliabilities of *emotional support* and *classroom organization*, and *instructional support* were high (.87–.90) across pre- and posttest. The domain averages were moderately to highly correlated within wave (rs = .64–.81, $p < .01$). Interrater reliability (IRR) was calculated using the 867 (32.7%) observations that were double-coded across pre- and posttest. IRR was calculated using a one-way random intraclass correlation (ICC). ICCs fell in the good to excellent range (.60–.93) for all CLASS dimension and domain scores across pre- and posttest (Cicchetti, 1994).

Social validity assessment. To examine participants' perceptions of the social importance and acceptability of the CARE for Teachers program, participants completed the CARE Acceptability Questionnaire. Participants who attended the final booster session completed the form at the end of the training day along with their self-assessment. Those who did not attend the booster session received an online version of the survey via e-mail. This measure was expanded to 23 items from its original 10-item version used in previous research (Jennings et al., 2013). Participants rated their overall satisfaction with the program and specific components (program content, facilitator skill, program length, setting, program design, communication from facilitators and coaching calls) on a 5-point scale (1 = *highly unsatisfied* to 5 = *highly satisfied*). They also rated their agreement to a set of statements related to perceived changes in teaching effectiveness and stress (1 = *strongly disagree* to 5 = *strongly agree*), perceived effects on students' behavior and academic performance (1 = *much worse* to 5 = *much better*), and perceived impact on job performance in comparison to other professional development programs (1 = *much lower* to 5 = *much higher*).

Results

In this section we first report preliminary analyses including the distributional properties of our sample, our handling of attrition and missing data, the comparability of intervention and control groups, our data reduction approach to teacher self-report outcomes, and results from our training process evaluation. This is followed by a description of our primary outcome analysis strategy and results of the impacts of CARE for Teachers on teachers and classroom interaction quality.

Preliminary Analyses

Descriptive and distributional properties of sample. We first examined distributions, outliers, multicollinearity, homogeneity of variance, and unusual patterns of missing data. Results revealed no statistically significant deviations from normality, variance, or multicollinearity on any scale variables. No unusual missing item patterns were detected, and, as described above, all standardized alpha values at baseline were in the acceptable range (α = .67–0.96).

Attrition and missing data. A variety of strategies were used to minimize attrition and total attrition levels were low (n = 15; 7%). Examination of possible intervention by attrition interactions yielded no statistically significant differences on pretest variables. Missing data were handled using the full information maximum likelihood estimation method under the assumption that missing is at random (Little & Rubin, 2002).

Comparability of intervention and control groups. Table 1 summarizes teacher- and classroom-level descriptive statistics by intervention and control status. Teacher-level descriptive statistics

Table 1
Teacher and Classroom Characteristics by Intervention and Control Status

Teacher and classroom characteristics	Total			Intervention			Control		
	Valid n	%	M (SD)	Valid n	%	M (SD)	Valid n	%	M (SD)
Cohort	224			118			106		
Cohort 1		23.7			22.0			25.0	
Cohort 2		76.3			78.0			75.0	
Teacher race/ethnicity	224			118			106		
White		33.4			34.8			31.1	
Non-White		66.6			65.2			68.9	
Classroom grade level	221			116			105		
Grade K–3		67.0			62.1			72.4	
Grade 4–5		33.0			37.9			27.6	
Classroom type	224			118			106		
General ed		84.4			83.1			85.9	
Other		15.7			16.9			14.1	
Student:teacher ratio	224		17.89 (5.26)	118		17.81 (5.38)	106		17.99 (5.16)
Proportion of IEP students	224		.10 (.09)	118		.10 (.09)	106		.10 (.09)
Proportion of suspended	205		.03 (.07)	107		.04 (.08)	98		.02 (.04)
Avg. learning support at home	214		3.56 (.53)	112		3.58 (.49)	102		3.53 (.57)

Note. Student:teacher ratio is an average of the number of students and teachers in each classroom at the time observations occurred. Proportion of IEP or 504 plan students and proportion of suspended collected from teachers except for 17 cases where these data were missing and therefore replaced with data from the New York City Department of Education Records. Avg. learning at home data collected from teacher report. See Measures section for more information.

include cohort and race/ethnicity. Classroom-level statistics include grade level, classroom type, student-teacher ratio, proportion of students with an IEP or 504 plan, proportion of students ever suspended, and teacher report of students' average learning support at home. The analyses found no statistically significant differences in baseline demographic characteristics between the two conditions. There were also no differences between groups on baseline outcome measures even after controlling for multiple pairwise contrasts. Thus, at baseline, randomization was effective in ensuring intervention and control groups were well balanced.

Data reduction of teacher self-report outcomes. To reduce the number of statistical tests across numerous teacher self-report assessments to the most theoretically and empirically relevant underlying constructs, we examined scale-level correlations and conducted exploratory and confirmatory factor analyses to identify a core set of meaningful higher-order constructs. In these analyses, we excluded teachers' physical symptoms and medication usage because they could not be grouped meaningfully into any categories. We utilized exploratory factor analysis using maximum likelihood estimation with promax oblique rotation to extract a set of cohesive factor constructs (Lorenzo-Seva & ten Berge, 2006).[4]

The first factor we identified as *mindfulness*, included all five mindfulness subscales from the FFMQ: observing, describing, acting with awareness, nonjudgmental and nonreactive; also included was the interpersonal mindfulness from the MTS. The second factor was labeled *psychological distress*, consisted of measures for depression (PHQ), anxiety (GAD), negative affect (PANAS), sleep disturbance (PROMIS), emotional exhaustion (MBI), and perceived stress (PPS). The third factor we identified as *time urgency* consisted of all subscales from the Time Urgency Scale: eating-related hurry, speech-related hurry, general hurry, task-related hurry, and competitiveness. We had originally included the TUS as a measure of psychological distress; however only task-related hurry cross-loaded on factor 2. The fourth factor, *teaching efficacy*, consisted of measures of teacher-reported self-efficacy in student engagement, instruction, and classroom management from the TSES.

We then subjected each derived factor to a confirmatory factor analysis to ensure adequate fit to our empirically derived measurement model. Examination of relative (CFI and TLI) and absolute overall model fit indices (RMSEA) suggested adequate fit of our measurement model to the data. Cronbach's alphas were .68 for *mindfulness*, .62 for *psychological distress*, .70 for *time urgency*, and .84 for *teaching efficacy*.

Three relevant measures—the cognitive reappraisal and expression suppression subscales of the ERQ and positive affect from the PANAS—were included in the initial exploratory factor analyses but did not load on any of the four empirically derived and theoretically consistent factors. In order to assess the construct of adaptive emotion regulation, assessed using cognitive reappraisal and expression suppression, we created a factor derived from averaging these measure items (after reverse scoring the expression suppression items). Cronbach's alpha was .67 for *adaptive emotion regulation*.

Social validity assessment. On the end-of-training CARE Acceptability Questionnaire, intervention teachers reported high levels of satisfaction (M = 4.47, SD = .50) with the program. Teachers also reported a high level of self-perceived improvement (M = 4.00, SD = .49). Specifically, teachers reported improvements in their well-being (88%) and self-awareness (96%) and many (63%) also indicated feeling less job stress as a result of the program. They also strongly agreed or agreed that as a result of CARE for Teachers they were "better able to promote awareness and concentration among their students" (87%), "manage classroom behaviors effectively and compassionately" (86%) and "better able to establish and maintain supportive relationships" with their students (91%).

Participants also reported seeing improvements in their students (M = 3.87, SD = .56). Specifically, teachers reported that their students were better or much better in regard to their pro-social behavior (78%), on-task behavior (75%), and academic performance (58%). Finally, teachers also were very willing to recommend the CARE program for other teachers (M = 4.44, SD = .57) Almost all teachers (95%) reported that they strongly agreed or agreed that this type of program should be integrated into preparation and in-service training.

Main Analyses

Outcome analysis strategy. Primary study outcomes were analyzed using two-level Hierarchical Linear Models for continuous outcomes or two-level Hierarchical Generalized Linear Models for count outcomes (e.g., physical symptoms and medication use) to account for the clustering of teachers within schools. For each model, only intercepts were allowed to vary randomly across schools. All analyses were performed in MPLUS, Version 7.2 (Muthén & Muthén, 1998-2012) using maximum likelihood estimation. We examined intervention impact on each outcome controlling for a set of covariates to maximize statistical power and precision of an intervention effect estimate. For self-report models, in the absence of baseline differences on demographic and baseline measures and/or preexisting hypotheses, only pretest scores and cohort were included. For classroom observational models additional covariates with known relationships to variations in teacher performance were included (grade level, classroom type, student-teacher ratio, teacher race, proportion of students with an IEP or 504 plan, proportion of students ever suspended, and teacher perceived average level of support for learning in the home). Pretest scores, student-teacher ratio, and average level of learning support at home were grand-mean centered. Effect sizes for statistically significant effects from the self-report and classroom observation models were calculated by dividing the adjusted mean difference by the unadjusted pooled standard deviation (Cohen, 1988). As recommended by the What Works Clearinghouse (Institute of Education Sciences, 2014), an improvement index (U3) was computed by calculating the difference between the percentile rank of the average teacher or classroom in the intervention condition and that of the average teacher or classroom in the control group.

As teacher medication usage and physical symptoms used count models, we first examined zero-count distributions and tested for

[4] A complete description of the factor analytic procedures and results can be found in the online supplementary materials. We have also compared the factor loading patterns between promax oblique rotation and varimax orthogonal rotation of the selected four-factor EFA model. The two factor loading patterns were consistent with a congruence coefficient = .95 (≥.90 is regarded as satisfactory similarity between two patterns; Lorenzo-Seva & ten Berge, 2006).

overdispersion using the overdispersion parameter, *alpha*. For medication usage and ache-related symptoms, the overdispersion parameters were not significantly different from zero ($ln(alpha) = 0.00$, $p > .99$ for both variables), satisfying the assumption of the Poisson distribution that the conditional mean and variance are equal (Long & Freese, 2006). We thus used a Poisson model for these outcome variables. As analyses of gastrointestinal-related symptoms did reveal significant levels of overdispersion ($ln(alpha) = 1.77$, $p < .05$), suggesting a departure from the Poisson distributional assumption, we used a Negative Binomial model, which corrects for overdispersion by adding a parameter that allows the conditional variance to be different from the conditional mean (Long & Freese, 2006).

Impact on teachers and classroom interaction quality. Below we report impact estimates of CARE for Teachers on teachers' self-report measures and observed classroom interaction quality. Table 2 presents means and standard deviations for teacher scales and classroom processes by intervention status.

Impact on teachers' self-report measures. Table 3 presents the results of the program impact on the five factors: four aggregate factors (i.e., teaching efficacy, mindfulness, psychological distress and time urgency); and the factor assessing teachers' adaptive emotion regulation. Statistically significant direct effects of CARE for Teachers were found for four out of the five factors. Participation in intervention led to statistically significant increases in adaptive emotion regulation, $t = 2.98$, $p = .005$ and mindfulness, $t = 2.71$, $p = .007$ and statistically significant reductions in psychological distress, $t = -1.99$, $p = .047$, and time urgency, $t = -2.32$, $p = .020$. The adjusted mean differences of 0.24 and 0.14 for adaptive emotion regulation and mindfulness correspond to effect sizes of 0.35 and 0.28, respectively. The adjusted mean differences of −0.13 and −0.10 for psychological distress and time urgency correspond to effect sizes of −0.18 and −0.20, respectively. There were no statistically significant effects on the factor assessing teaching efficacy.

The results of intervention impact on teachers' physical symptoms and medication use are displayed in Table 4. The program impact was not statistically significant for any of the three count outcomes, ache-related symptoms, gastrointestinal symptoms, and sum of medication; however, there was a tendency for CARE teachers to report fewer symptoms and medication use. The estimated incident rate ratios associated with the intervention were .805, .604, and .866 for ache-related symptoms, gastrointestinal symptoms, and medication usage, respectively. Effect sizes, measured as percentage change in expected count and calculated by subtracting 1 from incident rate ratio estimates and multiplying 100, indicated that CARE for Teachers reduced teachers' ache-related symptoms, gastrointestinal symptoms, and medication use by 19.5%, 39.6%, and 13.4%, respectively.

Impacts on classroom quality of interactions. Table 5 presents the results of the program impact on the CLASS domains and dimensions. The intervention had a statistically significant positive effect on the domain of emotional support ($t = 1.96$, $p = .051$, $ES = 0.22$), and positive effects on two of its associated dimensions, positive climate ($t = 2.15$, $p = .031$, $ES = 0.23$) and teacher sensitivity ($t = 1.99$, $p = .046$, $ES = 0.23$). There was also a marginally statistically significant positive effect of intervention on the domain of classroom organization ($t = 1.68$, $p = .093$, $ES = 0.19$) and a statistically significant positive effect on one of its associated dimensions, productivity ($t = 1.94$, $p = .052$, $ES = 0.23$). There was no statistically or marginally significant effect on the domain of instructional support or the associated dimensions.

Table 2
Teacher Scales and Classroom Processes by Intervention and Control Status

Teacher scales and classroom processes	Pre *M* (*SD*)		Post *M* (*SD*)	
	Intervention	Control	Intervention	Control
Teacher aggregate factors				
Adaptive emotion regulation	4.85 (0.70)	4.81 (0.71)	5.00 (0.70)	4.75 (0.68)
Teaching efficacy	7.15 (0.94)	7.01 (1.03)	7.31 (0.93)	7.22 (0.98)
Mindfulness	3.55 (0.43)	3.55 (0.42)	3.68 (0.49)	3.56 (0.46)
Psychological distress	2.57 (0.73)	2.67 (0.76)	2.37 (0.71)	2.51 (0.70)
Time urgency	3.24 (0.53)	3.37 (0.53)	3.16 (0.50)	3.31 (0.49)
Teacher physical distress				
Ache-related symptoms	1.27 (1.27)	1.17 (1.25)	.97 (1.11)	1.11 (1.18)
Gastrointestinal symptoms	0.29 (0.74)	.36 (0.72)	.21 (0.58)	0.37 (0.81)
Medication use	1.03 (1.10)	1.22 (1.10)	1.00 (0.92)	1.18 (1.19)
Quality of classroom interactions				
Emotional support	4.92 (0.80)	5.00 (0.70)	4.92 (0.76)	4.81 (0.74)
Positive climate	4.78 (1.10)	4.86 (1.01)	4.61 (1.02)	4.45 (0.98)
Negative climate	6.40 (0.70)	6.48 (0.60)	6.57 (0.56)	6.50 (0.58)
Teacher sensitivity	4.77 (0.97)	4.87 (0.89)	4.83 (0.98)	4.67 (0.98)
Respect for student perspective	3.73 (0.92)	3.81 (0.83)	3.69 (0.91)	3.64 (0.85)
Classroom organization	4.86 (0.90)	4.97 (0.80)	5.13 (0.86)	5.01 (0.88)
Behavior management	5.06 (1.06)	5.09 (0.88)	5.30 (1.02)	5.20 (0.99)
Productivity	5.13 (0.95)	5.28 (0.89)	5.45 (0.93)	5.26 (0.97)
Instructional learning formats	4.41 (0.97)	4.53 (0.93)	4.64 (0.86)	4.56 (0.93)
Instructional support	2.75 (0.67)	2.77 (0.71)	2.49 (0.65)	2.51 (0.65)
Concept development	2.38 (0.63)	2.54 (0.74)	2.18 (0.62)	2.25 (0.63)
Quality of feedback	3.03 (0.85)	3.01 (0.87)	2.82 (0.86)	2.76 (0.77)
Language modeling	2.83 (0.77)	2.76 (0.70)	2.47 (0.69)	2.53 (0.72)

Table 3
CARE for Teachers Impacts on Aggregate Factors

Aggregate factors	Estimate	*SE*	*t*	*p*	Effect size	U3	Improvement index %
Adaptive emotion regulation	.22	.08	2.98	.005*	.35	.64	13.68
Teaching efficacy	.07	.11	0.59	.556	.07	.53	2.79
Mindfulness	.13	.05	2.71	.007*	.28	.61	11.03
Psychological distress	−.13	.06	−1.99	.047*	−.18	.43	−7.14
Time urgency	−.10	.04	−2.32	.020*	−.20	.42	−7.93

* $p < .05$.

Post hoc analysis of subscales. For the three of four aggregate factors that showed statistically significant intervention impacts, we explored which subscales contributed to the overall effects (see Table 6).

Statistically significant program effects were found for 2 out of 6 subscales of mindfulness factor, nonjudging ($t = 2.04$, $p = .041$, Effect Size [*ES*] = 0.21) and observing ($t = 3.46$, $p = .001$, $ES = 0.41$); 2 out of 6 subscales of psychological distress, sleep ($t = 2.25$, $p = .024$, $ES = 0.26$) and emotional exhaustion ($t = -2.08$, $p = .037$, $ES = -0.22$); and, 2 out of 5 subscales of time urgency factor, speech ($t = -2.21$, $p = .027$, $ES = -0.18$) and task-related hurry ($t = -2.07$, $p = .038$, $ES = -0.22$).

Although we hypothesized that the intervention would have positive direct impacts on positive affect, as assessed using the PANAS positive affect subscale, it did not load with any conceptually appropriate aggregate factor. We therefore examined the program's direct impact on PANAS positive affect post hoc but found no statistically significant program effect.

Discussion

A growing body of evidence has demonstrated that teaching is a highly stressful profession and teacher stress has negative impacts on the quality of their classroom learning environment. Despite this evidence, little research has addressed ways to reduce teacher stress. The current study responded to this need by examining the efficacy of the CARE for Teachers program. The program was developed to promote the teacher social and emotional competencies described in Jennings and Greenberg's (2009) prosocial classroom model, proposing that when teachers lack certain social and emotional competences, their well-being erodes leading to a deterioration of the classroom climate and teacher stress. In contrast, teachers with high levels of social and emotional competencies are able to promote high quality classroom interactions that promote student learning.

The CARE for Teachers program elements of emotion skills instruction, mindful awareness and stress reduction, caring and listening practices were hypothesized to result in increases in adaptive emotion regulation, teaching efficacy and mindfulness and reductions in psychological and physical distress, as well as improvements in classroom interactions that promote learning (e.g., emotional support and classroom organization). In this discussion, we examine the practical importance of the impacts of CARE for Teachers on teacher and classroom outcomes and place these results within the context of the larger field of MBIs for teachers, including a review of study strengths and limitations, suggestions for future research and study implications.

Table 4
CARE for Teachers Impacts on Teacher Physical Distress

Impact	Estimate	*SE*	*t*	*p*	Effect size
Ache-related symptoms	−.22	.14	−1.59	.112	−19.5%
Gastrointestinal symptoms	−.50	.35	−1.46	.145	−39.6%
Medication use	−.14	.13	−1.08	.280	−13.4%

Practical Importance of Study Impacts

Here we review the study results and examine their practical importance in terms of the improvement index (What Works Clearinghouse; Institute of Education Sciences, 2014), and in relation to previous work.

Impact on teachers. Estimates of program impacts indicate that compared with control teachers, teachers who received CARE for Teachers reported significantly higher levels of functioning on four of the five factors that assessed broad domains hypothesized to be effected by the intervention. Compared with teachers in the control group, at the end of one school year intervention teachers showed higher levels of adaptive emotion regulation and mindfulness and lower levels of psychological distress and time urgency. These intervention effects were modest. In terms of the practical importance, on average, intervention teachers reported a 14% improvement in their ability to regulate their emotions ($U3 = 0.64$), an 11% increase in their overall mindfulness ($U3 = 0.61$), a 7% reduction in their reported psychological distress ($U3 = 0.43$), and 8% reduction in their sense of time urgency ($U3 = 0.42$) as compared with controls (see Table 3). These findings replicate previous work that has shown significant positive effects on similar outcomes (Crain et al., 2016; Flook et al., 2013; Jennings et al., 2013; Kemeny et al., 2012; Roeser et al., 2013; Taylor et al., 2016a, 2016b).

In addition to examining effects on the five broad domains of teacher-reported functioning, post hoc analyses on the psychological distress factor showed significant intervention effects on sleep disturbances (10% reduction; $U3 = 0.60$) and emotional exhaustion (9% reduction; $U3 = 0.41$; see Table 6). These results align with the results of the SMART program that found improvements in sleep and mood (Crain et al., 2016). Sleep problems have been negatively associated with well-being, job performance, and mental and physical health (Kuppermann et al., 1995). Emotional exhaustion, one dimension of occupational burnout (Maslach, Jackson, & Leiter, 1997), has also been negatively related to job performance, workplace satisfaction, teaching efficacy, and turnover (Collie, Shapka, & Perry, 2012; Klassen & Chiu, 2010; Wright & Cropanzano, 1998).

Table 5
CARE for Teachers Impacts on Quality of Classroom Interactions

Quality of classroom interactions	Estimate	*SE*	*t*	*p*	Effect size	U3	Improvement index %
Emotional support	.17	.08	1.96	.051*	.22	.59	8.71
Positive climate	.23	.11	2.15	.031*	.23	.59	9.10
Negative climate	.10	.06	1.53	.125	.17	.57	6.75
Teacher sensitivity	.23	.12	1.99	.046*	.23	.59	9.10
Respect for student perspective	.07	.11	0.67	.502	.08	.53	3.19
Classroom organization	.17	.10	1.68	.093	.19	.58	7.53
Behavior management	.13	.12	1.13	.258	.13	.55	5.17
Productivity	.22	.11	1.94	.052*	.23	.59	9.10
Instructional learning formats	.13	.11	1.23	.218	.14	.56	5.57
Instructional support	.00	.08	−0.03	.974	.00	.50	0.00
Concept development	−.03	.08	−0.36	.178	−.05	.48	−1.99
Quality of feedback	.07	.10	0.71	.478	.08	.53	3.19
Language modeling	−.07	.09	−0.80	.425	−.10	.46	−3.98

* $p < .05$.

Post hoc analyses revealed that the positive impacts on the broad construct of time urgency were due to significantly lower levels of speech- and task-related hurry. Intervention teachers reported a 7% reduction on the subscale of speech-related hurry (U3 = 0.43) and a 9% reduction in task-related hurry (*U3* = 0.41). The CARE for Teachers program applies mindful awareness practices to help teachers slow down their behavioral and thought patterns to gain a more realistic view of the time they have available for certain lessons and academic goals and to prioritize and plan accordingly. A reduction in time pressure may also lead to reporting less stress and exhaustion.

Intervention teachers reported a substantial 14% improvement (*U3* = 0.64) in adaptive emotion regulation compared with controls. This finding aligns with research that identified improved emotion regulation as a key to preventing teacher stress (Montgomery & Rupp, 2005). Adaptive emotion regulation involves both the ability to closely examine situations in which teachers experience difficult emotions and to be able to engage in cognitive reappraisal as well as to less often suppress their emotional expression. This finding is particularly important because CARE for Teachers specifically instructs teachers in how to recognize the physical sensations associated with the onset of emotion reactivity and to use mindful awareness practices and cognitive reappraisal to improve emotional self-regulation in the context of classroom. Emotion expression suppression has been shown to increase stress and impair well-being (Gross, 2002). It appears that CARE for Teachers supports teachers to use more adaptive ways of regulating, expressing, and coping with difficult emotions in the classroom.

Post hoc analyses of the intervention effects on mindfulness indicated significant improvements in the observing and nonjudging subscales of the mindfulness factor. CARE for Teachers par-

Table 6
CARE for Teachers Impacts on Subscales Within Aggregate Factors Showing Significant Effects

Subscales within aggregate factors	Estimate	*SE*	*t*	*p*	Effect size	U3	Improvement index %
Mindfulness	.13	.05	2.71	.007*	.28	.61	11.03
Describing	.10	.07	1.42	.155	.15	.56	5.96
Nonjudging	.17	.08	2.04	.041*	.21	.58	8.32
Awareness	.06	.07	0.83	.409	.08	.53	3.19
Observing	.29	.08	3.46	.001*	.41	.66	15.91
Nonreactive	.09	.08	1.11	.267	.15	.56	5.96
Interpersonal mindfulness	.10	.06	1.65	.100	.19	.58	7.53
Psychological distress	−.13	.06	−1.99	.047*	−.18	.43	−7.14
Depression	−.04	.06	−0.67	.503	−.07	.47	−2.79
Anxiety	−.10	.08	−1.15	.249	−.13	.45	−5.17
Negative affect	−.13	.09	−1.52	.130	−.16	.44	−6.36
Sleep disturbance	.24	.11	2.25	.024*	.26	.60	10.26
Emotional exhaustion	−.32	.15	−2.08	.037*	−.22	.41	−8.71
Perceived stress	−.17	.09	−1.82	.070	−.22	.41	−8.71
Time urgency	−.10	.04	−2.32	.020*	−.20	.42	−7.93
Hurried eating	−.09	.08	−1.06	.290	−.10	.46	−3.98
Speech-related hurry	−.14	.06	−2.21	.027*	−.18	.43	−7.14
General hurry	−.05	.10	−0.49	.627	−.05	.48	−1.99
Task-related hurry	−.14	.07	−2.07	.038*	−.22	.41	−8.71
Competitiveness	−.10	.06	−1.74	.082	−.16	.44	−6.36

* $p < .05$.

ticipants reported a substantial 16% improvement on the observing scale ($U3$ = 0.66) and an 8% improvement on the nonjudging subscale ($U3$ = 0.58). Together, these two dimensions of mindfulness may be particularly important for teachers. When a teacher can observe internal and external experiences with a nonjudgmental attitude, he or she may be better prepared to respond to classroom situations without making maladaptive attributions to events (e.g., perceiving student misbehavior as a personal affront). In this way, increases in mindfulness may support teachers' ability to reappraise emotionally provocative situations, reduce or prevent overreactions and feelings of burnout (Chang, 2009) and promote supportive classroom interactions (Roeser, 2016; Roeser, Skinner, Beers, & Jennings, 2012; Skinner & Beers, 2016).

It was somewhat surprising given the above findings that CARE teachers did not report higher levels of teaching efficacy compared with control teachers. It should be noted that in a previous study, CARE for Teachers demonstrated significant effects on teaching efficacy (Jennings et al., 2013). One factor that might explain this lack of replication is that baseline scores on teaching efficacy in the current sample were approximately one standard deviation higher than scores among teachers in the prior sample; thus, ceiling effects may have limited our capacity to detect significant intervention effects in this study.

Impact on classrooms. Compared with control teachers, intervention teachers provided higher levels of emotional support as observed by independent raters using the CLASS. Again, although significant, the effect was modest. On average, the intervention participants' CLASS scores improved by 9% on emotional support ($U3$ = 0.59; see Table 5). Within the emotional support domain, the performance dimensions of positive climate and teacher sensitivity both improved by 9% from pre to post ($U3$ = 0.59).

As reported by other investigators (Rivers, Brackett, Reyes, Elbertson, & Salovey, 2013), we found that teachers randomly assigned to the control group showed declines in emotional support from the beginning to the end of the school year. In contrast, the intervention showed a protective effect against this decline, with teachers trained in CARE for Teachers showing stable levels of classroom emotional support from pretest to posttest. The intervention showed similar protective effects for the positive climate dimension of emotional support, which reflects teachers' warmth, closeness and respect for students. The improvements in teacher's social and emotional competences may have contributed to this protective effect. When teachers experience less psychological distress, they are more likely to express positive emotions like smiling and laughter which promotes a supportive, positive climate (Pianta et al., 2008). In contrast to the above protective effects, the teacher sensitivity dimension of emotional support demonstrated statistically significant increases in the intervention group at post compared with a decline among controls. Teacher sensitivity reflects teachers' awareness and responsiveness to students' needs. Mindfulness, particularly the observing and nonjudging dimensions, may improve a teacher's ability to notice and respond to students' needs with more patience and understanding.

Although marginally statistically significant, an intervention effect was found on the domain of classroom organization, as evidenced by an 8% improvement ($U3$ = 0.57), with a statistically significant gain on the dimension of productivity (9% improvement, $U3$ = 0.59). Productivity represents how smoothly the classroom runs and how teachers maximize learning time. Improvements in productivity may result from the decreased time pressure CARE teachers reported. When teachers feel less pressure to meet daily and weekly goals, they may be better prepared and implement lesson plans effectively. Although these results are similar to those of a pilot study that showed improvements in classroom organization (Flook et al., 2013), the results here include a larger sample of teachers and use of more rigorous methods.

These findings are notable in that they are the first to demonstrate improvements in classroom interactions as a result of intervention efforts that do not explicitly focus on teachers' classroom management and instruction skills. In contrast to one widely used teacher coaching program that focuses on developing teacher's interactional skills with students based on CLASS dimensions (My Teaching Partner; Allen, Pianta, Gregory, Mikami, & Lun, 2011), CARE for Teachers primarily targets teachers' own social and emotional competencies through emotion skills instruction and mindful awareness practices. Explicit instruction in ways to promote teachers' emotional supportiveness and classroom organization is not part of the CARE for Teachers curriculum. However, following the prosocial classroom model improvements on these CLASS outcomes were hypothesized to follow from improvements in aspects of teacher social and emotional competences.

These demonstrated improvements at both the teacher and classroom levels provide support for key components of the CARE for Teachers logic model and the prosocial classroom model (Jennings & Greenberg, 2009) described above. They are also consistent with the theoretical model proposed by Roeser et al. (2012), Roeser (2016), and Skinner and Beers (2016) wherein mindfulness training promotes improved emotion regulation and coping which then leads to reductions in stress, burnout and distress, and increased energy and self-regulatory resources that can be invested in improving classroom interactions that support student learning.

Contextualization of Current Study Within Existing Evidence Base

Although the results of the current study are promising, it is useful to situate these findings within the context of prior research on teacher mindfulness. The present study has some notable methodological differences compared with prior studies. First, the present trial included a sample of 224 teachers which is substantially larger than any prior randomized trials examining the efficacy of teacher mindfulness programs conducted by Beshai et al. (2016; n = 89), Flook et al. (2013, n = 18), Franco et al. (2010; n = 36), Frank et al. (2015; n = 68), Poulin et al. (2008; n = 44), Taylor et al., 2016a, 2016b; n = 56), and Jennings et al. (2013, n = 50). Although these pilot investigations are critical for determining the feasibility and parameters of larger scale trial designs, estimates of effect size (d) become more precise in larger sample sizes (Leon, Davis, & Kraemer, 2011). As such, the present study provides a unique contribution to the literature in terms of the size, diversity, and scope of the population studied.

Second, the present study provided in-service to educators working in public school settings, as compared with teacher mindfulness studies that have examined outcomes for teacher trainees (Hue & Lau, 2015), teacher-assistant dyads (Gold et al., 2010), or parent-teacher dyads (Benn et al., 2012). Therefore, our study

results generalize to the common configuration of teacher-led elementary classrooms in diverse inner city settings.

Among randomized trials of a similar size and focus, the trial of SMART trial is most directly comparable (Roeser et al., 2013). Although both studies utilized similar measures, measurement strategies differed in important ways. For example, in the present study, emotional exhaustion, one factor of burnout, loaded on a broader aggregate factor of psychological distress, whereas the Roeser study combined the three subscales of the Maslach Burnout Inventory into a global measure of burnout. Roeser et al. (2013) found the SMART program significantly decreased teacher-reported levels of occupational stress ($d = -0.57$), burnout ($d = -0.76$), anxiety ($d = -0.71$), and depression ($d = -1.06$), whereas we found smaller effects for the emotional exhaustion component of burnout ($d = -0.22$) and no differences in anxiety or depression, despite significant differences in the aggregate of teacher psychological distress ($d = -0.18$). There may be several possible explanations for these somewhat discrepant findings. In terms of study sample characteristics, Roeser et al. (2013) utilized a sample of 113 elementary and secondary school teachers from Canada and the United States, whereas the current study sampled exclusively from inner city elementary level teachers in the United States. Although both studies had high levels of female participation (93% CARE for Teachers vs. 88% SMART), the present study sample had substantially more racial diversity (33% White) as compared with the Roeser et al. (2013) Canadian (67% White) and U.S. sample (93% White).

Aside from variations in sample demographic characteristics, other features related to the nature of implementation and measurement are other possible explanations for differences in observed outcomes. For example, in the present study CARE for Teachers was delivered during five in-service training days (30 contact hours) spread out across the entire school year, with baseline data collection occurring in fall and posttest collection in spring. In contrast, the SMART program implemented by Roeser et al. (2013) occurred during an 8-week period (11 afterschool sessions, 36 contact hours) during the spring semester, with baseline data collection occurring in February–March and posttest data collection in June. Although the total contact hours of these programs were quite similar, they differed with regards to the format (in-service vs. afterschool), number of sessions (five vs. 11), session duration (6 hr vs. 3–4 hr), and the timing of sessions during the school year (five sessions across the whole year vs. eight weeks during one semester only).

Study Strengths

The present study marks a promising step forward in the evaluation of MBIs for teachers. Among its strengths, it is the largest randomized controlled trial of a MBI for teachers to date and also the first to use both a randomized experimental design and accompanying analytic strategy that accounted for the clustering of teachers/classrooms within schools. Second, the trial showed effects on both teacher-self reports as well as independently observed outcomes and thus support the potential veracity of teacher reports. It is also the first rigorous trial of a MBI designed for teachers to demonstrate positive impacts on key aspects of the observed quality of classroom interactions. As such, it is the first demonstration that a MBI can have direct impacts on distal contextual factors that reflect positive social interactions. Another strength of the study is that the sample of teachers is racially and ethnically diverse (66.6% non-White) and the sample of classrooms observed covers the entire span of the elementary school grades (K-5).

Limitations

The present study had several limitations. The sample of schools and teachers participated in the study and the CARE for Teachers program voluntarily. For this reason, the results of the present study might not be generalizable to a sample of teachers mandated to participate in the program. Although CARE for Teachers demonstrated direct effects on four of five hypothesized teacher self-report factors and important dimensions of classroom interaction quality, these effects are small to moderate in magnitude. Another limitation is that this report only examined pre- and postintervention changes. It is likely that reductions in teachers' psychological distress and improvements in teachers' social and emotional competence and the quality of classroom interactions may change over time. Improvements could fade away in the absence of the intervention, or they may be augmented as teachers have more time to further develop their mindfulness and emotion skills to integrate them more comprehensively into their teaching practices and daily lives. The SMART program found continued improvement in mindfulness and occupational self-compassion and reductions in occupational stress, burnout, anxiety, and depression symptoms at 3-month follow-up (Roeser et al., 2013). Another report involving the same sample found continued reductions in bad mood at the 3-month follow up but also improvement in sleep quality which was not found immediately postintervention (Crain et al., 2016).

Suggestions for Future Research

Although the present study represents a promising advancement in the evaluation of MBIs for teachers and MBIs more generally, there is a need for further research to gain a more comprehensive understanding of the impact of MBIs on teacher, classroom and student outcomes. Understanding how geographic locale, grade level, and racial diversity may moderate the effectiveness of MBIs for teachers is an important area for future research. Furthermore, CARE for Teachers was delivered over 30 hr across 5 days and it will be important to study how variations in the intensity and duration of the program may be related to teacher outcomes. Reducing the time and intensity of the program, if findings were still positive, may affect the likelihood of school's adopting CARE for Teachers as an ongoing professional development program for all teachers within a school.

The improvements in dimensions in classroom interactions suggest that CARE for Teachers may improve student academic and behavioral outcomes not addressed in the present study but hypothesized in the prosocial classroom model (Jennings & Greenberg, 2009). Previous research has demonstrated that similar improvements in classroom emotional supportiveness and organization result in improvements in student-teacher relationships and student academic (Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008) and behavioral outcomes (Hamre & Pianta, 2005; Hoglund et al., 2015). Future research should examine student outcomes in relation to improvements in teacher and classroom outcomes.

MBIs for teachers may be most useful as a complement to social and emotional learning programs for students. Despite the modest effects found in the present study, such effect sizes are not uncommon in educational research (Hill, Bloom, Black, & Lipsey, 2008; Wilson, Lipsey, & Derzon, 2003). Combining such programs may have synergistic effects that boost the impacts of both programs.

Finally, cost–benefit analyses of programs such as CARE for Teachers and SMART could add to such programs' perceived value among school leaders and policymakers. Previous work has shown that MBIs may impact physical health in adult populations. In future studies, physiological measures (cortisol, blood pressure, immune function, etc.) could be assessed to examine effects on underlying physiological systems. In addition, it would be useful to assess teacher's health care utilization through study of insurance records as positive findings related to reductions in health care costs would be notable to school leaders and education policymakers.

Study Implications

Teacher stress and burnout is a critical issue in today's educational landscape, and only limited attention in policy and teacher training programs has been given to the matter (Greenberg et al., 2016). The results suggest that efforts to foster teachers' social and emotional competences may have significant impacts on both the cost and quality of education. In the long run, reducing teacher stress and burnout may reduce costs associated with teacher absenteeism, turnover, and health care, as well as lead to gains in classroom interaction quality and supportive teacher-student relationships that promote student positive social and emotional and academic development. The present study demonstrated CARE for Teachers to be a socially valid and well-received professional development program that can support the aforementioned goals.

In conclusion, this study provides the most rigorous evidence to date for the efficacy of a MBI to increase teacher social and emotional competence and the quality of classroom interactions. Additional research is needed to investigate whether this program shows longer-term effects on teachers and whether it is scalable to whole school or district-wide implementation.

References

Aber, J. L., Brown, J. L., & Jones, S. M. (2003). Developmental trajectories toward violence in middle childhood: Course, demographic differences, and response to school-based intervention. *Developmental Psychology, 39,* 324–348. http://dx.doi.org/10.1037/0012-1649.39.2.324

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011, August 19). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333,* 1034–1037. http://dx.doi.org/10.1126/science.1207998

Alliance for Excellent Education. (2014). *On the path to equity: Improving the effectiveness of beginning teachers* [Press release]. Retrieved from http://all4ed.org/press/teacher-attrition-costs-united-states-up-to-2-2-billion-annually-says-new-alliance-report/

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment, 13,* 27–45. http://dx.doi.org/10.1177/1073191105283504

Benn, R., Akiva, T., Arel, S., & Roeser, R. W. (2012). Mindfulness training effects for parents and educators of children with special needs. *Developmental Psychology, 48,* 1476–1487. http://dx.doi.org/10.1037/a0027537

Beshai, S., McAlpine, L., Weare, K., & Kuyken, W. (2016). A non-randomised feasibility trial assessing the efficacy of a mindfulness-based intervention for teachers to reduce stress and improve well-being. *Mindfulness, 7,* 198–208. http://dx.doi.org/10.1007/s12671-015-0436-1

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., . . . Devins, G. (2004). Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice, 11,* 230–241. http://dx.doi.org/10.1093/clipsy.bph077

Blitstein, J. L., Hannan, P. J., Murray, D. M., & Shadish, W. R. (2005). Increasing the degrees of freedom in existing group randomized trials through the use of external estimates of intraclass correlation: The DF* approach. *Evaluation Review, 29,* 241–267. http://dx.doi.org/10.1177/0193841X04273257

Brown, J. L., Jones, S. M., LaRusso, M. D., & Aber, J. L. (2010). Improving classroom quality: Teacher influences and experimental impacts of the 4Rs program. *Journal of Educational Psychology, 102,* 153–167. http://dx.doi.org/10.1037/a0018160

Buysse, D. J., Yu, L., Moul, D. E., Germain, A., Stover, A., Dodds, N. E., . . . Pilkonis, P. A. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep, 33,* 781–792.

Carson, R. L., Weiss, H. M., & Templin, T. J. (2010). Ecological momentary assessment: A research method for studying the daily lives of teachers. *International Journal of Research & Method in Education, 33,* 165–182. http://dx.doi.org/10.1080/1743727X.2010.484548

Chang, M. L. (2009). An appraisal perspective of teacher burnout: Examining the emotional work of teachers. *Educational Psychology Review, 21,* 193–218. http://dx.doi.org/10.1007/s10648-009-9106-y

Chang, M. L. (2013). Toward a theoretical model to understand teacher emotions and teacher burnout in the context of student misbehavior: Appraisal, regulation and coping. *Motivation and Emotion, 37,* 799–817. http://dx.doi.org/10.1007/s11031-012-9335-0

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290. http://dx.doi.org/10.1037/1040-3590.6.4.284

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. http://dx.doi.org/10.1177/001316446002000104

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24,* 385–396. http://dx.doi.org/10.2307/2136404

Collie, R. J., Shapka, J. D., & Perry, N. E. (2012). School climate and social-emotional learning: Predicting teacher stress, job satisfaction, and teaching efficacy. *Journal of Educational Psychology, 104,* 1189–1204. http://dx.doi.org/10.1037/a0029356

Collomp, K., Baillot, A., Forget, H., Coquerel, A., Rieth, N., & Vibarel-Rebot, N. (2016). Altered diurnal pattern of steroid hormones in relation to various behaviors, external factors and pathologies: A review. *Physiology & Behavior, 164,* 68–85.

Corcoran, K. M., Farb, N., Anderson, A., & Segal, Z. V. (2010). Mindfulness and emotion regulation: Outcomes and possible mediating mechanisms. In A. M. Kring & D. M. Sloan (Eds.), *Emotion regulation and psychopathology: A transdiagnostic approach to etiology and treatment* (pp. 339–355). New York, NY: Guilford Press.

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology, 108,* 100–102.

Crain, T. L., Schonert-Reichl, K. A., & Roeser, R. W. (2016). Cultivating teacher mindfulness: Effects of a randomized controlled trial on work,

home, and sleep outcomes. *Journal of Occupational Health Psychology.* Advance online publication. http://dx.doi.org/10.1037/ocp0000043

DeWeese, A., Jennings, P., Brown, J., Doyle, S., Davis, R., Rasheed, D., . . . Greenburg, M. (in press). Coding semi-structured interviews: Examining coaching calls within the CARE for Teachers Program. *Sage research methods cases.* Thousand Oaks, CA: Sage.

Domitrovich, C. E., Gest, S. D., Gill, S., Bierman, K. L., Welsh, J., & Jones, D. (2009). Fostering high-quality teaching with an enriched curriculum and professional development support: The Head Start REDI Program. *American Educational Research Journal, 46,* 567–597. http://dx.doi.org/10.3102/0002831208328089

Doyle, S. L., Jennings, P. A., DeWeese, A., & Frank, J. L. (2014, May). *Evaluating the fidelity of the cultivating awareness and resilience in education (CARE) Program.* Poster presented at the Society for Prevention Research Annual Meeting, Washington, DC.

Dworkin, A. G., & Tobe, P. F. (2014). The effects of standards-based school accountability on teacher burnout and trust relationships: A longitudinal analysis. In D. Van Maele, P. B. Forsyth, & M. Van Houtte (Eds.), *Trust and school life* (pp. 121–143). New York, NY: Springer Science + Business Media. http://dx.doi.org/10.1007/978-94-017-8014-8_6

Eberth, J., & Sedlmeier, P. (2012). The effects of mindfulness meditation: A meta-analysis. *Mindfulness, 3,* 174–189. http://dx.doi.org/10.1007/s12671-012-0101-x

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36,* 103–112. http://dx.doi.org/10.1207/S15326985EP3602_5

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272–299. http://dx.doi.org/10.1037/1082-989X.4.3.272

Felver, J. C., & Jennings, P. A. (2016). Introduction to the special issue: Applications of mindfulness-based interventions in school settings. *Mindfulness, 7,* 1–14. http://dx.doi.org/10.1007/s12671-015-0478-4

Flook, L., Goldberg, S. B., Pinger, L., Bonus, K., & Davidson, R. J. (2013). Mindfulness for teachers: A pilot study to assess effects on stress, burnout and teaching efficacy. *Mind, Brain and Education, 7,* 182–195. http://dx.doi.org/10.1111/mbe.12026

Franco, C., Mañas, I., Cangas, A. J., Moreno, E., & Gallego, J. (2010). Reducing teachers' psychological distress through a mindfulness training program. *The Spanish Journal of Psychology, 13,* 655–666. http://dx.doi.org/10.1017/S1138741600002328

Frank, J. L., Jennings, P. A., & Greenberg, M. T. (2016). Validation of the mindfulness in teaching scale. *Mindfulness, 7,* 155–163. http://dx.doi.org/10.1007/s12671-015-0461-0

Frank, J. L., Reibel, D., Broderick, P., Cantrell, T., & Metz, S. (2015). The effectiveness of Mindfulness-Based Stress Reduction on educator stress and well-being: Results from a pilot study. *Mindfulness, 6,* 208–216. http://dx.doi.org/10.1007/s12671-013-0246-2

Gallup. (2014). *State of America's schools: A path to winning again in education.* Washington, DC: Author. Retrieved from http://www.gallup.com/services/178709/state-america-schools-report.aspx?ays=n

Gold, E., Smith, A., Hopper, I., Herne, D., Tansey, G., & Hulland, C. (2010). Mindfulness-based stress reduction (MBSR) for primary school teachers. *Journal of Child and Family Studies, 19,* 184–189. http://dx.doi.org/10.1007/s10826-009-9344-0

Greenberg, M. T., Brown, J. L., & Abenavoli, R. M. (2016). *Teacher stress and health effects on teachers, students, and schools* [Issue brief]. Retrieved from http://www.rwjf.org/content/dam/farm/reports/issue_briefs/2016/rwjf430428

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology, 39,* 281–291. http://dx.doi.org/10.1017/S0048577201393198

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85,* 348–362. http://dx.doi.org/10.1037/0022-3514.85.2.348

Gu, Q., & Day, C. (2007). Teachers' resilience: A necessary condition for effectiveness. *Teaching and Teacher Education, 23,* 1302–1316. http://dx.doi.org/10.1016/j.tate.2006.06.006

Hagelskamp, C., Brackett, M. A., Rivers, S. E., & Salovey, P. (2013). Improving classroom quality with the RULER approach to social and emotional learning: Proximal and distal outcomes. *American Journal of Community Psychology, 51,* 530–543. http://dx.doi.org/10.1007/s10464-013-9570-x

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76,* 949–967. http://dx.doi.org/10.1111/j.1467-8624.2005.00889.x

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2,* 172–177. http://dx.doi.org/10.1111/j.1750-8606.2008.00061.x

Hoglund, W. L. G., Klingle, K. E., & Hosan, N. E. (2015). Classroom risks and resources: Teacher burnout, classroom quality and children's adjustment in high needs elementary schools. *Journal of School Psychology, 53,* 337–357. http://dx.doi.org/10.1016/j.jsp.2015.06.002

Hölzel, B. K., Carmody, J., Vangel, M., Congleton, C., Yerramsetti, S. M., Gard, T., & Lazar, S. W. (2011). Mindfulness practice leads to increases in regional brain gray matter density. *Psychiatry Research: Neuroimaging, 191,* 36–43. http://dx.doi.org/10.1016/j.pscychresns.2010.08.006

Hölzel, B. K., Hoge, E. A., Greve, D. N., Gard, T., Creswell, J. D., Brown, K. W., . . . Lazar, S. W. (2013). Neural mechanisms of symptom improvements in generalized anxiety disorder following mindfulness training. *NeuroImage Clinical, 2,* 448–458. http://dx.doi.org/10.1016/j.nicl.2013.03.011

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Hue, M. T., & Lau, N. S. (2015). Promoting well-being and preventing burnout in teacher education: A pilot study of a mindfulness-based programme for pre-service teachers in Hong Kong. *Teacher Development, 19,* 381–401. http://dx.doi.org/10.1080/13664530.2015.1049748

Institute of Education Sciences. (2014). *What Works Clearinghouse procedures and standards handbook* (Version 3.0). Washington, DC: Author. Retrieved from http://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

Iowa State University Extension and Outreach. (2010). *Strengthening families program: For parents and youth ages 10–14 facilitator delivery ratings.* Retrieved from http://www.episcenter.psu.edu/ebp/strengthening/sfpfidelityforms

Jennings, P. A. (2016a). CARE for Teachers: A mindfulness-based approach to promoting teachers' well-being and improving performance. In K. Schonert-Reichl & R. Roeser (Eds.), *The handbook of mindfulness in education: Emerging theory, research, and programs* (pp. 133–148). New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4939-3506-2_9

Jennings, P. A. (2016b). Mindfulness-based programs and the American public school system: Recommendations for best practices to ensure secularity. *Mindfulness, 7,* 176–178. http://dx.doi.org/10.1007/s12671-015-0477-5

Jennings, P. A., Frank, J. L., Snowberg, K. E., Coccia, M. A., & Greenberg, M. T. (2013). Improving classroom learning environments by Cultivating Awareness and Resilience in Education (CARE): Results of a randomized controlled trial. *School Psychology Quarterly, 28,* 374–390. http://dx.doi.org/10.1037/spq0000035

Jennings, P. A., & Greenberg, M. T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research, 79,* 491–525. http://dx.doi.org/10.3102/0034654308325693

Jennings, P. A., Snowberg, K. E., Coccia, M. A., & Greenberg, M. T. (2011). Improving classroom learning environments by Cultivating Awareness and Resilience in Education (CARE): Results of two pilot studies. *Journal of Classroom Interaction, 46,* 27–48.

Kabat-Zinn, J. (1982). An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: Theoretical considerations and preliminary results. *General Hospital Psychiatry, 4,* 33–47. http://dx.doi.org/10.1016/0163-8343(82)90026-3

Kabat-Zinn, J. (2003). Mindfulness-based interventions in context: Past, present, and future. *Clinical Psychology: Science and Practice, 10,* 144–156. http://dx.doi.org/10.1093/clipsy.bpg016

Kavanagh, D. J., & Bower, G. H. (1985). Mood and self-efficacy: Impact of joy and sadness on perceived capabilities. *Cognitive Therapy and Research, 9,* 507–525. http://dx.doi.org/10.1007/BF01173005

Kemeny, M. E., Foltz, C., Cavanagh, J. F., Cullen, M., Giese-Davis, J., Jennings, P., . . . Ekman, P. (2012). Contemplative/emotion training reduces negative emotional behavior and promotes prosocial responses. *Emotion, 12,* 338–350. http://dx.doi.org/10.1037/a0026118

Khoury, B., Lecomte, T., Fortin, G., Masse, M., Therien, P., Bouchard, V., . . . Hofmann, S. G. (2013). Mindfulness-based therapy: A comprehensive meta-analysis. *Clinical Psychology Review, 33,* 763–771. http://dx.doi.org/10.1016/j.cpr.2013.05.005

Klassen, R. M., & Chiu, M. M. (2010). Effects of teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology, 102,* 741–756. http://dx.doi.org/10.1037/a0019237

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders, 114,* 163–173. http://dx.doi.org/10.1016/j.jad.2008.06.026

Kuppermann, M., Lubeck, D. P., Mazonson, P. D., Patrick, D. L., Stewart, A. L., Buesching, D. P., & Filer, S. K. (1995). Sleep problems and their correlates in a working population. *Journal of General Internal Medicine, 10,* 25–32. http://dx.doi.org/10.1007/BF02599573

Kyriacou, C. (2011). Teacher stress: From prevalence to resilience. In J. Langan-Fox & C. L. Cooper (Eds.), *Handbook of stress in the occupations* (pp. 161–173). Northampton, MA: Edward Elgar Publishing. http://dx.doi.org/10.4337/9780857931153.00027

Landy, F. J., Rastegary, H., Thayer, J., & Colvin, C. (1991). Time urgency: The construct and its measurement. *Journal of Applied Psychology, 76,* 644–657. http://dx.doi.org/10.1037/0021-9010.76.5.644

Larsen, R. J., & Kasimatis, M. (1991). Day-to-day physical symptoms: Individual differences in the occurrence, duration, and emotional concomitants of minor daily illnesses. *Journal of Personality, 59,* 387–423. http://dx.doi.org/10.1111/j.1467-6494.1991.tb00254.x

Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research, 45,* 626–629. http://dx.doi.org/10.1016/j.jpsychires.2010.10.008

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley. http://dx.doi.org/10.1002/9781119013563

Long, S. J., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2nd ed.). College Station, TX: Stata Press.

Lorenzo-Seva, U., & ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2,* 57–64. http://dx.doi.org/10.1027/1614-2241.2.2.57

Markow, D., Macia, L., & Lee, H. (2013). *The MetLife survey of the American teacher: Challenges for school leadership.* New York, NY: Metropolitan Life Insurance Company.

Maslach, C., Jackson, S. E., & Leiter, M. P. (1997). Maslach burnout inventory. In C. P. Zalaquett & R. J. Wood (Eds.), *Evaluating stress: A book of resources* (pp. 191–218). Lanham, MD: Scarecrow.

Milkie, M. A., & Warner, C. H. (2011). Classroom learning environments and the mental health of first grade children. *Journal of Health and Social Behavior, 52,* 4–22. http://dx.doi.org/10.1177/0022146510394952

Montgomery, C., & Rupp, A. A. (2005). A meta-analysis for exploring the diverse causes and effects of stress in teachers. *Canadian Journal of Education/Revue Canadienne Education, 28,* 458–486.

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

New York City Independent Budget Office. (2014, May). Demographics and work experience: A statistical portrait of New York City's public school teachers. *Schools brief.* Retrieved from http://www.ibo.nyc.ny.us/iboreports/2014teacherdemographics.pdf

Oberle, E., & Schonert-Reichl, K. A. (2016). Stress contagion in the classroom? The link between classroom teacher burnout and morning cortisol in elementary school students. *Social Science & Medicine, 159,* 30–37. http://dx.doi.org/10.1016/j.socscimed.2016.04.031

Osher, D., Sprague, J., Weissberg, R. P., Axelrod, J., Keenan, S., Kendziora, K., & Zins, J. E. (2007). A comprehensive approach to promoting social, emotional, and academic growth in contemporary schools. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., Vol. 5, pp. 1263–1278). Bethesda, MD: National Association of School Psychologists.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal, 45,* 365–397. http://dx.doi.org/10.3102/0002831207308230

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS) manual: K–3.* Baltimore, MD: Brookes.

Poulin, P. A., Mackenzie, C. S., Soloway, G., & Karayolas, E. (2008). Mindfulness training as an evidenced-based approach to reducing stress and promoting well-being among human services professionals. *International Journal of Health Promotion and Education, 46,* 72–80. http://dx.doi.org/10.1080/14635240.2008.10708132

Raudenbush, S., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29,* 5–29. http://dx.doi.org/10.3102/0162373707299460

Rhoads, C. H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics, 36,* 76–104. http://dx.doi.org/10.3102/1076998610379133

Richards, J. (2012). Teacher stress and coping strategies: A national snapshot. *The Educational Forum, 76,* 299–316. http://dx.doi.org/10.1080/00131725.2012.682837

Rivers, S. E., Brackett, M. A., Reyes, M. R., Elbertson, N. A., & Salovey, P. (2013). Improving the social and emotional climate of classrooms: A clustered randomized controlled trial testing the RULER Approach. *Prevention Science, 14,* 77–87. http://dx.doi.org/10.1007/s11121-012-0305-2

Roeser, R. W. (2016). Processes of teaching, learning, and transfer in mindfulness-based interventions (MBIs) for teachers: A contemplative educational perspective. In K. Schonert-Reichl & R. Roeser (Eds.), *The handbook of mindfulness in education: Emerging theory, research, and programs* (pp. 133–149). New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4939-3506-2_10

Roeser, R. W., Schonert-Reichl, K., Jha, A., Cullen, M., Wallace, L., Wilensky, R., . . . Harrison, J. (2013). Mindfulness training and reductions in teacher stress and burnout: Results from two randomized, waitlist-control field trials. *Journal of Educational Psychology, 105,* 787–804. http://dx.doi.org/10.1037/a0032093

Roeser, R. W., Skinner, E., Beers, J., & Jennings, P. A. (2012). Mindfulness training and teachers' professional development: An emerging area

of research and practice. *Child Development Perspectives, 6,* 167–173. http://dx.doi.org/10.1111/j.1750-8606.2012.00238.x

Schochet, P. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33,* 62–87. http://dx.doi.org/10.3102/1076998607302714

Sharma, M., & Rush, S. E. (2014). Mindfulness-based stress reduction as a stress management intervention for healthy individuals: A systematic review. *Journal of Evidence-Based Complementary & Alternative Medicine, 19,* 271–286. http://dx.doi.org/10.1177/2156587214543143

Skinner, E., & Beers, J. (2016). Mindfulness and teachers' coping in the classroom: A developmental model of teacher stress, coping, and everyday resilience. In K. Schonert-Reichl & R. Roeser (Eds.), *The handbook of mindfulness in education: Emerging theory, research, and programs* (pp. 88–118). New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4939-3506-2_7

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine, 166,* 1092–1097. http://dx.doi.org/10.1001/archinte.166.10.1092

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009). *Optimal design for longitudinal and multilevel research: Documentation for the "optimal design" software.* Retrieved from http://sitemaker.umich.edu/groupbased/files/od-manual-v200-20090722.pdf

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42,* 893–898. http://dx.doi.org/10.1016/j.paid.2006.09.017

Sutton, R. E., & Wheatley, K. F. (2003). Teachers' emotions and teaching: A review of the literature and directions for future research. *Educational Psychology Review, 15,* 327–358. http://dx.doi.org/10.1023/A:1026131715856

Tang, Y. Y., Hölzel, B. K., & Posner, M. I. (2015). The neuroscience of mindfulness meditation. *Nature Reviews Neuroscience, 16,* 213–225. http://dx.doi.org/10.1038/nrn3916

Taylor, C., Harrison, J., Haimovitz, K., Oberle, E., Thomson, K., Schonert-Reichl, K., & Roeser, R. W. (2016a). Erratum to "Examining ways that a mindfulness-based intervention reduces stress in public school teachers: A mixed-methods study." *Mindfulness, 7,* 1499. http://dx.doi.org/10.1007/s12671-016-0620-y

Taylor, C., Harrison, J., Haimovitz, K., Oberle, E., Thomson, K., Schonert-Reichl, K., & Roeser, R. W. (2016b). Examining ways that a mindfulness-based intervention reduces stress in public school teachers: A mixed-methods study. *Mindfulness, 7,* 115–129. http://dx.doi.org/10.1007/s12671-015-0425-4

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-Cultural Psychology, 38,* 227–242. http://dx.doi.org/10.1177/0022022106297301

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17,* 783–805. http://dx.doi.org/10.1016/S0742-051X(01)00036-1

Tsouloupas, C. N., Carson, R. L., Matthews, R., Grawitch, M. J., & Barber, L. K. (2010). Exploring the association between teachers' perceived student misbehaviour and emotional exhaustion: The importance of teacher efficacy beliefs and emotion regulation. *Educational Psychology, 30,* 173–189. http://dx.doi.org/10.1080/01443410903494460

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: U.S. Department of the Army.

Werthamer-Larsson, L. (1994). Methodological issues in school-based services research. *Journal of Clinical Child Psychology, 23,* 121–132. http://dx.doi.org/10.1207/s15374424jccp2302_2

Wethington, E. (2000). Contagion of stress. *Advances in Group Processes, 17,* 229–253. http://dx.doi.org/10.1016/S0882-6145(00)17010-9

Williams, J. M. G., & Kabat-Zinn, J. (2011). Mindfulness: Diverse perspectives on its meaning, origins, and multiple applications at the intersection of science and dharma. *Contemporary Buddhism, 12,* 1–18. http://dx.doi.org/10.1080/14639947.2011.564811

Wilson, S. J., Lipsey, M. W., & Derzon, J. H. (2003). The effects of school-based intervention programs on aggressive behavior: A meta-analysis. *Journal of Consulting and Clinical Psychology, 71,* 136–149. http://dx.doi.org/10.1037/0022-006X.71.1.136

Wright, T. A., & Cropanzano, R. (1998). Emotional exhaustion as a predictor of job performance and voluntary turnover. *Journal of Applied Psychology, 83,* 486–493. http://dx.doi.org/10.1037/0021-9010.83.3.486

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442. http://dx.doi.org/10.1037/0033-2909.99.3.432

Received September 2, 2015
Revision received December 7, 2016
Accepted December 13, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 7, 1029–1047

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000171

A Double-Edged Sword? On the Benefit, Detriment, and Net Effect of Dimensional Comparison on Self-Concept

Hanno Müller-Kalthoff
University of Kiel

Malte Jansen
Humboldt University of Berlin

Irene M. Schiefer
University of Bamberg

Friederike Helm, Nicole Nagy, and Jens Möller
University of Kiel

Dimensional comparison theory (DCT; Möller & Marsh, 2013) assumes that students compare their academic achievement intraindividually across domains to form domain-specific self-concepts. Upward dimensional comparisons are believed to lead to lower self-concepts in the worse-off domain, while downward dimensional comparisons should lead to higher self-concepts in the better-off domain. Furthermore, DCT assumes the net effect of upward and downward dimensional comparisons to be beneficial to the self. To test these assumptions, 3 experiments and 2 field studies were conducted investigating the relative effects of upward and downward dimensional comparisons as well as their net effect. In Studies 1 (N = 149), 2 (N = 150) and 3 (N = 300), participants were asked to infer self-concepts of fictitious students after receiving experimentally manipulated information about their achievements in 2 domains, whereas participants in Studies 4 (N = 2,268) and 5 (N = 20,662) assessed their own self-concepts in German and mathematics. In all studies, downward dimensional comparisons resulted in higher self-concepts, whereas upward dimensional comparisons led to lower self-concepts. The net effect of dimensional comparisons was always found to be not statistically different from zero. The findings therefore support the central prediction of DCT on the discreteness of the effects of upward and downward dimensional comparisons, yet do not support the assumed positivity of their net effect. Furthermore, results indicate the effect patterns to be rather universal as they were stable across different samples, domains, achievement situations, research designs, and types of assessment.

Keywords: dimensional comparison, net effect, other-ratings, self-concept, self-ratings

Academic self-concepts (Marsh, 2006; Marsh & Shavelson, 1985; Shavelson, Hubner, & Stanton, 1976) are based on a variety of comparison processes (e.g., Skaalvik & Skaalvik, 2002) comparing a certain target with a certain standard (Biernat & Eidelman, 2007; Mussweiler, 2003). For example, the internal/external frame of reference model (I/E model; Marsh, 1986; Möller, Müller-Kalthoff, Helm, Nagy, & Marsh, 2016) postulates that students form their self-concept in an academic domain by comparing their own performance (target) simultaneously to an external standard (e.g., the performance of their peers) as well as to an internal standard (e.g., their own performance in other domains). Whereas the use of an external reference is based on social comparison (Festinger, 1954), the use of an internal reference is based on dimensional comparison (Möller & Köller, 2001).

According to dimensional comparison theory (DCT; Möller, Helm, Müller-Kalthoff, Nagy, & Marsh, 2015; Möller & Marsh, 2013), dimensional comparisons occur when people compare their own performances intraindividually across domains (e.g., in school, the verbal and mathematical domain or school subjects like English and mathematics, respectively) and typically lead to contrast effects, or more precisely, negative path coefficients from achievement in one domain to self-concept in the other (Marsh, 1986; Möller, Pohlmann, Köller, & Marsh, 2009). For example, two students with similar achievements in English would develop different beliefs about their verbal ability depending on their performance in mathematics: A student comparing his English grade upward to a better mathematics grade would develop a lower verbal self-concept than a student comparing it downward to a worse mathematics grade.

In two diary studies, Möller and Husemann (2006) found that students use dimensional comparisons spontaneously in everyday life to maintain their mood states and self-worth by comparing their current state mainly between different academic domains, personal relationships, leisure time activities, and their own per-

This article was published Online First February 6, 2017.

Hanno Müller-Kalthoff, Department of Educational Psychology, University of Kiel; Malte Jansen, Institute for Educational Quality Improvement, Humboldt University of Berlin; Irene M. Schiefer, Department of Educational Research, University of Bamberg; Friederike Helm, Nicole Nagy, and Jens Möller, Department of Educational Psychology, University of Kiel.

Malte Jansen is now at the German Institute for International Educational Research (DIPF), Berlin, Germany.

Correspondence concerning this article should be addressed to Hanno Müller-Kalthoff, Department of Educational Psychology, University of Kiel, Olshausenstraße 75, 24118 Kiel, Germany. E-mail: hanno@mueller-kalthoff.com

sonality or characteristics. Furthermore, a substantial and growing body of path analytical research (for an overview, see Möller et al., 2009) as well as experimental studies (e.g., Dickhäuser, Seidler, & Kölzer, 2005; Möller & Köller, 2001; Möller & Savyon, 2003; Pohlmann & Möller, 2006, 2009; Tietjens & Niewerth, 2005) has provided support for the effects of dimensional comparisons in the past. In their meta-analysis of 69 path-analytical studies on the I/E model (N = 125,308), Möller et al. (2009) showed that the positive effects of downward dimensional comparisons as well as the negative effects of upward dimensional comparisons "are not restricted to a particular achievement or self-concept measure or to specific age groups, gender groups, or countries" (p. 1157). Moreover, the generalized internal/external frame of reference model (GI/E model; Möller et al., 2016) assumes dimensional comparisons "to be rather general comparison processes not only limited to the formation of academic self-concepts alone, but to apply to evaluations of different constructs as well" (p. 44; see also Möller et al., 2015; Möller & Marsh, 2013). In line with this assumption, a growing number of studies provides evidence of effects of dimensional comparisons on a variety of variables other than self-concepts, such as students' domain-specific academic emotions (e.g., Goetz, Frenzel, Hall, & Pekrun, 2008), interests (e.g., Schurtz, Pfost, Nagengast, & Artelt, 2014), intrinsic motivation (e.g., Marsh, Abduljabbar, et al., 2015), perceptions of their learning environment (e.g., Arens & Möller, 2016), and career choices (e.g., Parker et al., 2012) as well as peoples' perceptions of their honesty (e.g., Möller & Savyon, 2003) and their actual pro-social behavior (e.g., Brown & Smart, 1991). Furthermore, a handful of studies suggest that the cognitive process underlying dimensional comparisons is not limited to evaluations of one's own characteristics, but also apply to evaluations of others' characteristics (e.g., other-rated self-concepts; see Dickhäuser, 2005; Möller, 1999, 2005; Möller & Köller, 1997; Müller-Kalthoff, Helm, & Möller, 2015). Thus, the evaluative process underlying dimensional comparisons can be assumed to be highly generalizable (cf., Möller & Marsh, 2013; Möller et al., 2016).

As self-concepts have a long-lasting importance and propagating impact on domain-specific motivation (see the expectancy-value theory; Eccles et al., 1983; Wigfield & Eccles, 2000), subsequent academic achievements (e.g., Marsh & Craven, 2006; Retelsdorf, Köller, & Möller, 2014; Valentine, DuBois, & Cooper, 2004), self-concepts (e.g., Möller, Retelsdorf, Köller, & Marsh, 2011; Möller, Zimmermann, & Köller, 2014; Niepel, Brunner, & Preckel, 2014), and course-choices (e.g., Köller, Daniels, Schnabel, & Baumert, 2000; Nagy et al., 2008; Nagy, Trautwein, Baumert, Köller, & Garrett, 2006), the effects of dimensional comparisons gain further importance in the field of educational psychology as they may lead to biased self-views and significantly affect students' academic careers. In the present research, we therefore looked more closely at the rarely investigated interplay of dimensional upward and downward comparisons by comparing their respective effects on academic self-concept and by analyzing their net effect (i.e., the sum of their self-concept reducing and enhancing effects).

Upward and Downward Dimensional Comparisons

The negative path coefficients between the verbal and mathematical domains as described in the I/E model each comprise two separate, yet intertwined comparison processes (Pohlmann & Möller, 2009): Students comparing, for example, their mathematical achievement with their own worse-off verbal achievement (*downward dimensional comparison*) may enhance their mathematical self-concept, while students comparing their mathematical achievement with a better-off verbal achievement (*upward dimensional comparison*) may lower their mathematical self-concept (for a similar reasoning concerning social comparisons, see Möller & Pohlmann, 2010). As every upward comparison from target to standard is accompanied by a downward comparison from standard to target, Möller and Marsh (2013) conclude that "dimensional comparisons are a double-edged sword, as they lower the self-concept in the worse off domain while raising it in the better off domain" (p. 546). Furthermore, Möller et al. (2015) point out that dimensional comparisons "often lead to an over- or underestimation of own abilities" (p. 9) in the better-off or worse-off domain, respectively.

With regard to the motivational factors that may lead to and influence dimensional comparisons, Möller and Marsh (2013) assume that the negative effects of upward dimensional comparisons may serve a self-evaluative function as "accurate perceptions of one's strengths and weaknesses are critical for self-understanding and form the basis of informed life decisions" (p. 547) and "even negative effects of achievement in a better off domain on self-concept in a weaker domain seem reasonable, as they prevent wrong decisions" (p. 550). In addition, Möller et al. (2015) believe that self-differentiation motivation may also trigger dimensional comparisons as they lead to "the perception of highly faceted self-concepts with many distinct cognitive self-aspects" (p. 432) resulting in high self-complexity (cf., Linville, 1985). For example, a student having a differentiated view of his own strengths and weaknesses may use dimensional comparisons to facilitate important life decisions, like what hobbies to focus on or what career to pursue (e.g., Dickhäuser, Reuter, & Hilling, 2005; Nagy et al., 2006), even if that means that he or she is over- or underestimating his or her own abilities (cf., Möller et al., 2015).

On the other hand, Möller et al. (2015) suggest that the positive effect of downward dimensional comparisons may be indicative of self-serving information processing (e.g., Taylor & Brown, 1988), represented by motivations like self-maintenance, self-improvement and self-enhancement (cf., Tesser, 1988). More precisely, they hypothesize that students confronted with a threat to one of their own abilities may switch their focus to a better-off domain (cf., Baumeister, 1982; Baumeister & Jones, 1978; Boney-McCoy, Gibbons, & Gerrard, 1999; Dodgson & Wood, 1998; Steele, 1988). The resulting downward comparison from that better-off domain to the domain in jeopardy may then compensate for the overall threat to their selves by increasing their mood (Möller & Husemann, 2006) as well as their self-concept in the better-off domain (Pohlmann & Möller, 2009). For example, "when a student's self-worth is threatened by negative feedback for a target domain, their attention may switch to certain standard domains offering an escape from negative outcomes and an opportunity for compensatory self-enhancement" (Möller et al., 2015, p. 432).

> Furthermore, "it can be very helpful to switch attention from a suffering target domain to one's own particular strengths as a comparison standard, when trying to improve in an area of weakness

(although it leads to costs in self-concepts in the target domain)" (Möller et al., 2015, p. 432).

Unfortunately, the majority of path-analytical research on the I/E model (cf., Möller et al., 2009) does not provide information on the effect sizes of upward and downward dimensional comparisons (cf., Pohlmann & Möller, 2009). To date, in fact, only two field studies (i.e., Pohlmann & Möller, 2009, Studies 1 and 2) and three experimental studies (i.e., Möller & Köller, 2001, Study 2; Pohlmann & Möller, 2009, Study 3; Strickhouser & Zell, 2015, Study 1) have utilized research designs capable of disentangling the self-concept reducing effects of upward dimensional comparisons from the self-concept enhancing effects of downward dimensional comparisons. Specifically, these five studies examined the effect of achievement feedback in one domain (standard) on self-concept in another domain (target) by holding achievement in the target domain constant at an average level while varying achievement in the standard domain between below-average (downward dimensional comparison), average/no feedback (horizontal dimensional comparison), and above-average levels (upward dimensional comparison). This experimental manipulation allowed the researchers to hold the influence of social comparison constant on self-concept in the target domain, and also attribute self-concept changes in the target domain to dimensional comparisons only (i.e., the achievement variation in the standard domain). Whereas participants in the field studies were assigned to the achievement groups depending on their actual achievement in the target and standard domain, participants in the experimental studies received systematically manipulated performance feedback following the completion of two tasks. By using the horizontal dimensional comparison condition as a control group, these studies were also able to investigate and compare the relative effects of downward and upward dimensional comparisons on self-concept in the target domain. More importantly, by subtracting the negative effect of upward comparisons (d_{up}) from the positive effect of downward comparisons (d_{down}), they were also able to compute the so-called *net effect* of dimensional comparisons, that is, the sum of the self-concept reducing effects of upward dimensional comparisons and the self-concept enhancing effects of downward dimensional comparisons ($\Delta d = d_{up} + d_{down}$). The net effect of dimensional comparisons is not only a measure for whether dimensional comparisons in total lead to net gains or net losses in regard to the self (cf., Pohlmann & Möller, 2009), but may also serve as an indicator of the main motivation underlying the use of dimensional comparisons. For example, a significantly positive net effect might indicate self-serving rather than self-evaluative reasons to use dimensional comparisons when evaluating one's own abilities in a domain.

Interestingly, and despite using a similar design, results of these prior studies differed rather significantly (see Table 1). Three studies found statistically significant effects for downward dimensional comparisons on participants' self-concepts (Pohlmann & Möller, 2009, Studies 1 to 3), amounting to a positive net effect, two studies found a similar effect pattern for participants' satisfaction with their achievement (Möller & Köller, 2001, Study 2; Pohlmann & Möller, 2009, Study 3), and one study found only statistically significant effects for upward dimensional comparisons on participants' self-concepts (Möller & Köller, 2001, Study 2), amounting to a negative net effect. Unfortunately, none of these studies tested the net effect of dimensional comparison for statistical significance (or rather the difference of the effects of upward and downward dimensional comparisons, as suggested by Möller & Pohlmann, 2010). In fact, the only study that has reported this test (Strickhouser & Zell, 2015, Study 1) found the effects of upward and downward dimensional comparisons on self-concept as well as satisfaction to be equal in size and their net effect to be not significantly different from zero.

Table 1
Effect Sizes for Upward, Downward, as Well as the Net Effect of Dimensional Comparisons

Manuscript	Study	N	Domain	Variable	d_{up}	d_{down}	Δd
Möller & Köller (2001)	2	45	Math	Self-concept	−0.95	0.10	−0.85
				Satisfaction	−0.61	1.50	0.89
Pohlmann & Möller (2009)	1	93	Math	Self-concept	−0.46	0.66	0.20
		101	Verbal	Self-concept	−0.06	0.59	0.53
	2	458	Math	Self-concept	−0.20	0.37	0.17
	3	82	Math	Self-concept	−0.29	0.75	0.46
				Satisfaction	−0.12	0.99	0.87
Strickhouser & Zell (2015)	1	122	Verbal	Self-concept	−0.71	0.64	−0.07
				Satisfaction	−0.60	0.63	0.02
Present	1	149	Both	Self-concept[a]	−0.40	0.44	0.04
	2	150	Both	Self-concept[a]	−0.44	0.60	0.16
	3	300	N/A	Self-concept[a]	−0.63	0.79	0.16
	4	1,076.5	Verbal	Self-concept	−0.25	0.24	−0.01
		843.6	Math	Self-concept	−0.22	0.29	0.07
	5	9,263.4	Verbal	Self-concept	−0.29	0.39	0.09
		7,215.1	Math	Self-concept	−0.44	0.36	−0.08

Note. d_{up} = effect of upward dimensional comparison; d_{down} = effect of upward dimensional comparison; Δd = net effect of upward and downward dimensional comparisons (d_{up} + d_{down}). All effect sizes according to Cohen (1988). All studies reported qualify to estimate the net effect of dimensional comparisons by measuring the effect of upward and downward dimensional comparisons relative to a control group.
[a] Variables in Studies 1 to 3 of the present research were other-rated measures, whereas variables in all other studies were self-rated measures.

In light of the larger amount of empirical evidence for a beneficial net effect of dimensional comparisons (i.e., six out of nine results), Möller et al. (2015) proposed that "the net effect of dimensional comparisons is positive" (Hypothesis VIII; p. 434) and argue that "the overall benefit in the use of dimensional comparisons suggests that people use dimensional comparisons [primarily] as a way of regulating their self-evaluations in a self-enhancing way" (p. 434; see also Möller & Marsh, 2013). Accordingly, dimensional comparisons would be mostly driven by self-enhancement motivation and the positive effects of downward dimensional comparisons should be more prevalent and more pronounced than negative effects of upward dimensional comparisons. However, their hypothesis remains untested and may be a bit premature. In a meta-analysis of all previous results (see Table 1) using a mixed effects model, we found the average positive effect of downward dimensional comparisons, $d_{down} = 0.64$, $SE = 0.09$, $p < .001$, 95% CI [0.46, 0.82], to be not significantly larger than the average negative effect of upward dimensional comparisons, $d_{up} = -0.39$, $SE = 0.09$, $p < .001$, 95% CI [-0.57, -0.21], and their net effect in total to be not significantly different from zero, $\Delta d = 0.12$, $SE = 0.07$, $p = .059$, 95% CI [-0.01, 0.25], accordingly.[1] Therefore, previous empirical findings on the net effect of dimensional comparisons appear to be more heterogeneous than previously believed and do not support the hypothesis of Möller et al. (2015) that they sum up to a positive net effect on academic self-concept. Instead, the meta-analytic findings reported here support DCT's central assumption that both upward *and* downward dimensional comparisons yield meaningful effects, forming a "double-edged sword" with a net effect near zero.

Present Research

To resolve the conflicting previous findings on the occurrence and strength of the effects of upward and downward dimensional comparisons and to test the hypothesis of Möller et al. (2015) that the net effect of dimensional comparisons is positive, we replicated and extended the design used by Pohlmann and Möller (2009) in three experimental studies and two field studies. More precisely, we analyzed the effect of varying achievement levels in one domain (standard) on self-concept in another domain (target). To measure and compare self-concept differences, achievement in the target domain was controlled for and set to an average level (therefore also holding the effect of social comparison on self-concept in the target domain constant). Achievement in the standard domain was either above-average, average, or below-average leading to upward, horizontal, or downward dimensional comparisons from the perspective of the target domain's average achievement, respectively. Therefore, our design gave us the opportunity to not only measure self-concept differences in the target domain following upward and downward dimensional comparisons relative to a horizontal comparison control group, but to compute and test their net effect for statistical significance.

In line with DCT, we predicted upward dimensional comparisons (i.e., lower achievement in the target than in the standard domain) to reduce self-concept in the target domain and downward dimensional comparisons (i.e., better achievement in the target compared to the standard domain) to enhance self-concept. Furthermore, and in line with the assumptions of DCT, we expected the absolute effect size of downward dimensional comparisons to be larger than that of upward dimensional comparisons, resulting in a positive net effect of dimensional comparisons. Figure 1 illustrates the design and expected effects of dimensional comparisons in all five studies.

To investigate the evaluative process underlying dimensional comparisons, we examined both other-ratings and self-assessments of self-concept. In Studies 1 to 3, participants received an experimentally manipulated vignette portraying an achievement feedback situation in which a fictitious high-school student received grades in German and mathematics. Participants were asked to infer his self-concept in both domains. Whereas Study 1 tested our hypotheses on the effects of dimensional comparisons in a sample proximal to the vignette's academic context (German high-school students rating a fictitious high-school student's self-concepts), Study 2 served as a replication with a more distal sample (German university students rating a fictitious high-school student's self-concepts). In Study 3, the vignette was set in a nonschool setting describing a fictitious person receiving feedback in two standardized achievement tests. Finally, Studies 4 and 5 tested our hypotheses with more ecologically valid data from two large-scale field studies, examining the association between upward and downward dimensional comparisons and self-assessed verbal and mathematical self-concepts of 6th grade (Study 4) and 9th grade (Study 5) students. Overall, the five studies provide information on the generalizability of the evaluative process underlying dimensional comparisons and the stability of the effects of upward and downward dimensional comparisons as well as their net effect across different populations, achievement contexts, research designs (experimental vs. correlative), and types of assessment (other-ratings vs. self-assessments).

Study 1

Study 1 combines and extends the designs of previous experimental studies on dimensional comparisons (cf., Dickhäuser, 2005; Pohlmann & Möller, 2009). Similar to the design used by Dickhäuser (2005), participants read a vignette of a fictitious student receiving grades in German and mathematics and were then asked to infer his self-concept in both domains. Similar to the experimental design used by Pohlmann and Möller (2009, Study 3), the student's grades were systematically manipulated between participants to induce upward, horizontal, and downward dimensional comparisons (see Figure 1).

Previous studies investigating the net effect of dimensional comparisons (i.e., Möller & Köller, 2001, Study 2; Pohlmann & Möller, 2009, Study 3; Strickhouser & Zell, 2015, Study 1) gave participants manipulated achievement feedback and then used participants' self-assessed self-concept as the dependent variable. In Study 1, we manipulated the achievement feedback of a fictitious

[1] Absolute effect sizes of upward and downward dimensional comparisons did not differ in size, $\beta = -0.40$, $p = .053$, $\tau = .02$, and Q tests indicated individual effects to be homogeneous, upward: $Q_w(8) = 8.08$, $p = .425$; downward: $Q_w(8) = 11.77$, $p = .162$. Additional moderation analyses showed that the net effect of dimensional comparisons was not moderated by varying sample size, $\beta = -0.08$, $p = .799$, $\tau = .14$, the type of self-concept measure (evaluative vs. affective), $\beta = 0.47$, $p = .097$, $\tau = .10$, and the research design used (correlational vs. experimental), $\beta = -0.05$, $p = .881$, $\tau = .15$. However, because of the small number of studies available, results may not be reliable.

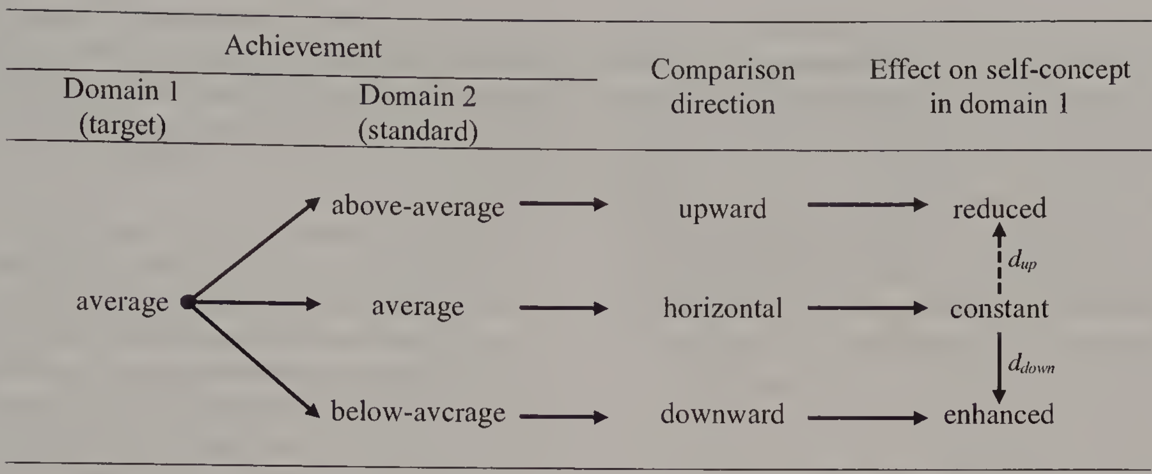


Figure 1. Expected relative effects of upward and downward dimensional comparisons on self-concept in domain 1 as a function of achievement in domain 1 (target) and domain 2 (standard). In line with DCT, the (enhancing) effect of downward dimensional comparisons (d_{down}) is assumed to be stronger than the (reducing) effect of upward dimensional comparisons (d_{up}; indicated by a dotted line) resulting in a positive net effect ($\Delta d = d_{up} + d_{down}$). Adapted from "On the Benefit of Dimensional Comparisons," by B. Pohlmann & J. Möller, 2009, *Journal of Educational Psychology, 101*, p. 250. Copyright 2009 by American Psychological Association. Reprinted/adapted with permission.

student, instead, and used his self-concept inferred by the participants as the dependent measure. Although self-assessed self-concepts and other-rated self-concepts are technically different constructs, the evaluative process underlying dimensional comparisons has been shown to be generalizable to a variety of settings and constructs and their effects to be similar (cf., Möller & Marsh, 2013; Möller et al., 2009, 2015, 2016). Furthermore, when trying to take the perspective of another, an observer has to overcome his or her own egocentrically biased knowledge and exercise cognitive control to adjust his own perspective toward the perspective of the observed actor (for an overview, see Epley & Caruso, 2009). This process often ends prematurely though, as observers tend to stop the adjusting process once a plausible estimate has been met (Epley, Keysar, Van Boven, & Gilovich, 2004, Study 5). As we asked participants in Study 1 to put themselves in the position of the fictitious student and gave them no other information except his grades (zero-acquaintance; e.g., Hirschmüller, Egloff, Nestler, & Back, 2013), participants should be forced to apply heuristics from their own experiences to the evaluative process. Ergo, if they believe self-concept reductions following upward dimensional comparisons and self-concept enhancement following downward dimensional comparisons to be plausible (e.g., from their own prior experiences), these beliefs should influence their judgment. In line with DCT and the GI/E model, we therefore assumed the effects of upward and downward dimensional comparisons to be similar as if participants evaluated their own abilities.

Method

Sample. The sample consisted of N = 149 German high-school students, 63.8% female, M = 17.54 years old, SD = 1.04, from Grades 11 to 13 of four secondary schools qualifying for university entrance ("Gymnasium") in Northern Germany. Participation was voluntary and parents gave informed consent prior to data collection.

Independent variable: Dimensional comparison direction. The direction of dimensional comparison served as the independent variable and was operationalized in three levels (upward vs. horizontal vs. downward dimensional comparison) by systematically varying the fictitious student's school grades in German and mathematics. Grades in Germany typically range from 1 ("*very good*") to 6 ("*poor*"), where a grade of 3 ("*satisfactory*") is considered to be average. While the grade in the first domain (target) was reported to be average (3, "*satisfactory*") in all experimental conditions, the fictitious student's achievement in the second domain (standard) was varied systematically between above-average (1, "*very good*"), average (3, "*satisfactory*"), and below-average (5, "*insufficient*") levels to induce upward, horizontal, and downward dimensional comparisons from the target to the standard domain, respectively.

Dependent Variables.

Self-concept. Academic self-concept in the target as well as the standard domain was used as the dependent variable and measured by five items adapted from Möller and Köller (2001), German: Cronbach's alpha = .91; Mathematics: Cronbach's alpha = .90. Participants were asked to answer the items from the fictitious student's point of view and responded on a 6-point scale from 0 ("*does not apply at all*") to 5 ("*applies completely*"). A sample item read: "*Although I certainly try hard, I find German/mathematics quite difficult.*" Negatively worded items were reverse coded and higher scores indicated higher self-concept.

Procedure. The study was conducted during classes and lasted about 15 minutes. Participants were randomly assigned to one of the three experimental conditions (upward vs. horizontal vs. downward dimensional comparison) and answered German paper-and-pencil questionnaires. The instruction read: "*A student receives the following grades in his mid-term school report.*" The instruction was followed by the fictitious student's grades in German and mathematics and then continued: "*Please put yourself in the position of this student, try to imagine how the following statements regarding the two subjects would apply to him from his point of view, and choose the respective answers for him.*" The school subject used as the target domain (German vs. mathematics)

was randomized between participants. Participants were then asked to infer the fictitious student's self-concepts in German and mathematics. Participants inferred higher self-concept in German than in mathematics across all experimental conditions, $F(1, 143) = 15.33$, $p < .001$, $\eta^2 = .01$, yet no interaction between domain and dimensional comparison direction was found to be significant, $F(2, 143) = 0.42$, $p = .658$. Therefore, the samples were combined for further analyses. After completing the questionnaire, participants answered questions about their demographical background, were thanked, and debriefed.

Results

Means and standard deviations for other-rated self-concepts by comparison condition are presented in Table 2.

Manipulation checks. To test whether the grades used in the present study were indeed perceived as different achievement levels, we examined the effect of the grade in the standard domain on self-concept in the standard domain. As the grade in the target domain was average in all experimental conditions, differences in self-concept in the standard domain should stem from our systematic variation of the grades in the standard domain only. In line with the positive path coefficients found for achievement in a domain on corresponding self-concept (cf., Möller & Pohlmann, 2010; Möller et al., 2009), we expected to find self-concept to increase with better grades. A one-factor analysis of variance (ANOVA) revealed a significant main effect of the grade in the standard domain (above-average vs. average vs. below-average) on self-concept in the standard domain, $F(2, 146) = 106.34$, $p < .001$, $\eta^2 = .12$. Post hoc Scheffé tests confirmed the hypothesized effect pattern: An above-average grade led to higher other-rated self-concepts than an average grade, $M_{dif} = 1.58$, $p < .001$, $d = 1.65$, 95% CI [1.19, 2.10], whereas a below-average grade led to lower other-rated self-concepts than an average grade, $M_{dif} = -1.27$, $p < .001$, $d = -1.22$, 95% CI [−1.65, −0.79]. This confirms that the manipulation of the different achievement levels was successful.

In addition, we examined the integrity of our horizontal comparison condition as a control group by testing if an average achievement (3, "*satisfactory*") had the same effect on self-concept in both the target and standard domain. A paired *t* test showed that self-concept in both domains did not differ significantly, $M_{dif} = 0.10$, $t(49) = 0.47$, $p = .642$, confirming that average achievement in the target and in the standard had the same effect on other-rated self-concept.

Pairwise comparisons. To test our main hypotheses on the effects of upward and downward dimensional comparisons as well as their net effect, three planned, nonorthogonal pairwise comparisons were conducted using independent-samples *t* tests. In line with recommendations made by Kirk (2013), a method by Holm (1979) was used to maintain a family wise error rate of $\alpha < .05$.

Upward dimensional comparisons. To test whether upward dimensional comparisons reduced self-concept in the target domain, we compared self-concept in the upward condition with self-concept in the horizontal condition. In line with our predictions, participants inferred significantly lower self-concept in the upward condition than in the horizontal condition, $M_{dif} = -0.38$, $t(146) = -2.01$, $p = .023$, $d_{up} = -0.40$, 95% CI [−0.79, 0.00].

Downward dimensional comparisons. To test whether downward dimensional comparisons enhanced self-concept in the target domain, we compared self-concept in the downward condition with self-concept in the horizontal condition. Again in line with our predictions, participants inferred significantly higher self-concept in the downward condition than in the horizontal condition, $M_{dif} = 0.44$, $t(146) = 2.29$, $p = .012$, $d_{down} = 0.44$, 95% CI [0.04, 0.84].

Net effect. Finally, to test whether the effects of upward and downward dimensional comparisons differed in size, we subtracted the difference between self-concepts in the upward and horizontal conditions from the difference between self-concepts in the downward and horizontal conditions and compared it to zero.[2] Note that this contrast not only tested whether the relative effects of upward and downward dimensional comparisons differ in size, but also whether their net effect differed from zero. In contrast to our expectations, we found no significant difference between the absolute values of the upward-horizontal difference and the downward-horizontal difference, $t(146) = 0.17$, $p = .433$. Hence, the relative effects of upward, $d_{up} = -0.40$, and downward dimensional comparisons, $d_{down} = 0.44$, were of equal size and their net effect, though marginally positive, was not statistically different from zero, $\Delta d = 0.04$, 95% CI [−0.36, 0.44].

Discussion

In Study 1, we investigated the effects of upward and downward dimensional comparisons by giving high-school students a vignette portraying an artificial achievement feedback situation and asking them to infer a fictitious student's self-concept. By experimentally varying the student's achievement in German and mathematics, we were able to experimentally trigger upward and downward dimensional comparisons (while holding the effect of social comparison constant), and disentangle their individual effects. Results were identical whether the target domain was German or mathematics: In line with our predictions derived from DCT, participants inferred higher self-concept in the target domain following downward dimensional comparisons, whereas they inferred lower self-concept following upward dimensional comparisons. In contrast to the assumptions of DCT, however, the reducing and enhancing effects of upward and downward dimensional comparisons were of nearly equal size. Their net effect was positive, but not significantly different from zero.

Study 1 is the first study showing statistically significant effects of *both* upward and downward dimensional comparisons. Our results therefore substantiate one of the central assumptions of DCT that dimensional comparisons are more than compensatory self-enhancement (cf., Baumeister & Jones, 1978; Sedikides & Gregg, 2008) as they do not only lead to enhancement in the better-off domain, but also to reduction in the worse-off domain. Moreover, our results also extend DCT significantly as they indicate the effects of upward and downward dimensional comparisons to be equally strong (see also Strickhouser & Zell, 2015).

In line with previous studies investigating the effect of dimensional comparisons on students' self-assessed self-concepts, Study 1 found effects of dimensional comparisons on other-rated self-concepts. Therefore, Study 1 provides further support for the

[2] The formula for this contrast was adapted from Möller and Pohlmann (2010): $t = (M_{\text{downward}} - M_{\text{horizontal}}) - (M_{\text{horizontal}} - M_{\text{upward}})/\text{SQRT}((1/n_{\text{downward}} + 4/n_{\text{horizontal}} + 1/n_{\text{upward}}) * MS_{\text{error}})$.

Table 2

Study 1 Means and Standard Deviations of Other-Rated Self-Concept by Domain (Target vs. Standard), Dimensional Comparison Direction (Upward vs. Horizontal vs. Downward), and Subject Used for the Target Domain (German vs. Mathematics)

		Target (Subject 1)					
		German		Mathematics		Overall	
Domain	Direction	*M* (*SD*)	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*n*
Target	Upward	2.55 (0.94)	25	2.16 (0.68)	25	2.36 (0.83)	50
	Horizontal	3.09 (1.00)	25	2.38 (1.03)	25	2.74 (1.07)	50
	Downward	3.49 (0.87)	25	2.84 (0.87)	24	3.17 (0.92)	49
Standard	Upward	4.30 (1.02)	25	4.53 (0.59)	25	4.41 (0.83)	50
	Horizontal	2.34 (1.01)	25	3.34 (0.89)	25	2.84 (1.07)	50
	Downward	1.58 (1.01)	25	1.56 (1.02)	24	1.57 (1.00)	49

Note. $N = 149$. If the subject used for the target domain was German, the subject used for the standard domain was mathematics and vice versa.

assumptions of DCT and the GI/E model that the evaluative process underlying dimensional comparisons is highly generalizable. Study 1's results were also in line with a recently published study on self-assessed self-concepts (see also Strickhouser & Zell, 2015) that also found the net effect of dimensional comparisons to be near zero, but differed from those of other studies on self-assessed self-concepts (Möller & Köller, 2001, Study 2; Pohlmann & Möller, 2009) that found the net effect to be either negative or positive. Hence, we cannot rule out the possibility that the motivational factors underlying dimensional comparisons are different for self-assessments and other-ratings and lead to different outcomes. To test this, we assessed participants' self-assessed self-concepts in two field studies, namely Studies 4 and 5.

Study 2

Study 2 used the same experimental design as Study 1 on university students instead of high-school students, thereby testing the validity of our initial findings in a different sample.

Method

Sample. The sample consisted of $N = 150$ German university students, 68.0% female, $M = 22.79$ years old, $SD = 2.68$, from different undergraduate and graduate programs in education science at the University of Kiel, Northern Germany. Participation was voluntary and parents gave informed consent prior to data collection.

Procedure. The study was conducted during lectures and lasted about 15 minutes. Participation was voluntary and participants received the same material and measures as those in Study 1, that is, they were asked to infer the self-concepts of a fictitious high-school student based on his grades in German and mathematics, German self-concept: Cronbach's alpha = .86; Mathematics self-concept: Cronbach's alpha = .89. Again, participants inferred higher self-concept in German than in mathematics across all experimental conditions, $F(1, 144) = 35.28$, $p < .001$, $\eta^2 = .01$, yet there was no significant interaction between domain and dimensional comparison direction, $F(2, 144) = 2.59$, $p = .078$.

Results

Means and standard deviations for other-rated self-concepts by comparison condition are presented in Table 3.

Manipulation checks. An ANOVA revealed a significant main effect of the grade in the standard domain (above-average vs. average vs. below-average) on self-concept in the standard domain, $F(2, 147) = 259.19$, $p < .001$, $\eta^2 = .16$. Post hoc Scheffé tests revealed that an above-average grade led to higher other-rated self-concepts

Table 3

Study 2 Means and Standard Deviations of Other-Rated Self-Concept by Domain (Target vs. Standard), Dimensional Comparison Direction (Upward vs. Horizontal vs. Downward), and Subject Used for the Target Domain (German vs. Mathematics)

		Target (Subject 1)					
		German		Mathematics		Overall	
Domain	Direction	*M* (*SD*)	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*n*
Target	Upward	2.54 (0.68)	25	1.90 (0.66)	25	2.22 (0.74)	50
	Horizontal	3.08 (0.81)	25	2.11 (0.82)	25	2.60 (0.94)	50
	Downward	3.23 (0.53)	25	2.88 (0.49)	25	3.06 (0.54)	50
Standard	Upward	4.53 (0.54)	25	4.62 (0.55)	25	4.58 (0.54)	50
	Horizontal	2.34 (0.90)	25	2.68 (0.63)	25	2.51 (0.78)	50
	Downward	1.42 (0.75)	25	1.66 (0.63)	25	1.54 (0.70)	50

Note. $N = 150$. If the subject used for the target domain was German, the subject used for the standard domain was mathematics and vice versa.

than an average grade, $M_{dif} = 2.06$, $p < .001$, $d = 3.07$, 95% CI [2.49, 3.65], whereas a below-average grade led to lower other-rated self-concepts than an average grade, $M_{dif} = -0.97$, $p < .001$, $d = -1.31$, 95% CI [−1.74, −0.88]. Furthermore, a paired t test showed that self-concept in both domains did not differ significantly in the horizontal condition, $M_{dif} = 0.08$, $t(49) = 0.54$, $p = .593$. Therefore, the experimental manipulation of the different achievement levels was successful.

Pairwise comparisons. To test our main hypotheses for Study 2, pairwise comparisons were conducted as described in Study 1.

Upward dimensional comparisons. Like in Study 1, participants inferred significantly lower self-concept in the upward condition than in the horizontal condition, $M_{dif} = -0.38$, $t(147) = -2.48$, $p = .007$, $d_{up} = -0.44$, 95% CI [−0.84, −0.05].

Downward dimensional comparisons. Participants also inferred significantly higher self-concept in the downward condition than in the horizontal condition, $M_{dif} = 0.46$, $t(147) = 3.04$, $p = .001$, $d_{down} = 0.60$, 95% CI [0.20, 1.00].

Net effect. In addition, we again found no significant difference, $t(147) = 0.32$, $p = .375$, between the absolute effects of upward, $d_{up} = -0.44$, and downward dimensional comparisons, $d_{down} = 0.60$, indicating them to be of equal size and their net effect not significantly different from zero, $\Delta d = 0.16$, 95% CI [−0.24, 0.56].

Discussion

In Study 2, we successfully replicated the results of Study 1 in a sample (university students) more distal to the achievement situation depicted in our experimental vignette (a fictitious student receiving grades in school). As in Study 1, other-rated self-concept was significantly lowered following upward dimensional comparisons and significantly enhanced following downward dimensional comparisons. Furthermore, the net effect of dimensional comparisons was also small and positive, yet still not significantly different from zero.

As in Study 1, the results of Study 2 supported our hypothesis that both upward and downward dimensional comparisons yield significant effects. They also showed that these effects generalize to other-ratings of self-concept and across different populations within the same research design. Furthermore, Study 2 substantiates Study 1 on the near-zero net effect of dimensional comparisons, which suggests again that the effects of upward and downward dimensional comparisons are equally strong.

Study 3

In Study 3, we aimed at testing whether the effects found in Studies 1 and 2 for German and mathematics are specific to these two subjects and the verbal and mathematical domains or if they apply similarly to different domains, as well. To this end, we extended the design used in Studies 1 and 2 by using a generic achievement situation. Instead of a high-school student receiving grades in school, the vignette in Study 3 depicted a nonschool achievement context with a fictitious person receiving feedback in two standardized tests, simply labeled Test 1 and Test 2, each measuring achievement in an unspecified achievement domain.

Method

Sample. The sample consisted of $N = 300$ German university students, 65.3% female, $M = 21.93$ years old, $SD = 2.83$. Participants were recruited by announcements in social community networks (e.g., Facebook) and were offered to enter a lottery for a jackpot totaling 200 € for compensation.

Independent variable: Dimensional comparison direction. The direction of dimensional comparison again served as the independent variable, but was this time operationalized by systematically varying the fictitious person's scores in both tests. The test scores were each interval-scaled ranging from 0 to 100 points. Similar to the design used in Studies 1 and 2, the test score for Test 1 (target) was reported to be in the middle of the scale ("*51 of 100 points*") in all experimental conditions. The test score for Test 2 (standard) was systematically varied, instead, to induce different dimensional comparison conditions: In the upward condition the score for Test 2 was high ("*87 of 100 points*"), in the horizontal condition it was in the middle of the scale ("*54 of 100 points*"), and in the downward condition it was low ("*15 of 100 points*"). The test scores used in the horizontal condition to represent medium achievement (*51 vs. 54*) were slightly varied to increase credibility. Furthermore, the scores used to represent high (*87*) and low (*15*) achievements were selected to each differ exactly 36 points from the achievement in the target (*51*).

Dependent variables.

Self-concept. We used the same items as in Studies 1 and 2 to measure self-concept in the target and the standard domain; both Cronbach's alpha > .84. Items were worded differently, though, to account for the vignette's different achievement context (i.e., Test 1 and Test 2 instead of German and mathematics).

Procedure. Participants answered an online questionnaire similar to the pen-and-paper versions used in Studies 1 and 2. Participants were randomly assigned to one of the three experimental conditions (upward vs. horizontal vs. downward dimensional comparison) and received the following instruction: "*A person completes two standardized achievement tests. The person receives the following scores in the two tests.*"[3] The instruction was followed by the person's achievement in two standardized achievement tests simply labeled Test 1 and Test 2. Similar to Studies 1 and 2, participants were then asked to infer his self-concept in Tests 1 and 2. After completing the questionnaire, participants answered questions about their demographical background, were debriefed, and finally forwarded to a separate web page to apply for the lottery.

Results

Means and standard deviations for other-rated self-concepts by comparison condition are presented in Table 4.

Manipulation checks. An ANOVA revealed a significant main effect of the test score in the standard domain (*87* vs. *54* vs.

[3] The instruction further read: "*The tests measure different/similar domains of performance that are important/not important to the person.*" The difference of the domains measured by the two tests (*different* vs. *similar*) as well as their importance to the fictitious person (*important* vs. *not important*) were also varied systematically between participants to account for possible effects and entered as control variables into our statistical design. As neither the experimental variation of domain difference, $F(1, 288) = 0.17$, $p = .682$, nor of importance, $F(1, 288) = 0.45$, $p = .501$, showed any significant influence on the dependent variable and no interaction effects were found, all $F < 1.26$, all $p > .285$, both control variables were excluded from further analyses and the samples combined.

Table 4
Study 3 Means and Standard Deviations of Other-Rated Self-Concept by Domain (Target vs. Standard) and Dimensional Comparison Direction (Upward vs. Horizontal vs. Downward)

Domain	Direction	*M* (*SD*)	*n*
Target	Upward	1.77 (0.81)	100
	Horizontal	2.34 (0.99)	100
	Downward	3.05 (0.78)	100
Standard	Upward	3.76 (0.74)	100
	Horizontal	2.53 (0.90)	100
	Downward	1.83 (0.92)	100

Note. $N = 300$.

15 points) on self-concept in the standard domain, $F(2, 297) = 130.67$, $p < .001$, $\eta^2 = .07$. Post hoc Scheffé tests revealed that high achievement led to higher other-rated self-concepts than medium achievement, $M_{dif} = 1.23$, $p < .001$, $d = 1.49$, 95% CI [1.18, 1.81], whereas low achievement led to lower other-rated self-concepts than medium achievement, $M_{dif} = -0.70$, $p < .001$, $d = -0.78$, 95% CI [−1.06, −0.49]. This confirms that the manipulation of the different achievement levels was again successful.

Unexpectedly, a paired *t* test showed that self-concept in both domains differed significantly in the horizontal condition, $M_{dif} = 0.19$, $t(99) = 3.43$, $p < .001$, $d = 0.20$, 95% CI [−0.08, 0.48]. Hence, the test scores given for medium achievement in the target domain (*51 points*) led to lower other-rated self-concept than the medium achievement given in the standard domain (*54 points*) despite their small mathematical difference (*3 points*).

Pairwise comparisons. Again, the same pairwise comparisons were conducted as described in Studies 1 and 2.

Upward dimensional comparisons. Like in Studies 1 and 2, participants inferred significantly lower self-concept in the upward condition than in the horizontal condition, $M_{dif} = -0.57$, $t(297) = -4.63$, $p < .001$, $d_{up} = -0.63$, 95% CI [−0.91, −0.35].

Downward dimensional comparisons. Participants also inferred significantly higher self-concept in the downward condition than in the horizontal condition, $M_{dif} = 0.71$, $t(297) = 5.77$, $p < .001$, $d_{down} = 0.79$, 95% CI [0.50, 1.08].

Net effect. Finally, the absolute effect sizes of upward, $d_{up} = -0.63$, and downward dimensional comparisons, $d_{down} = 0.79$, were not significantly different in size, $t(297) = 0.66$, $p = .256$, and their net effect, though again small and positive, was not significantly different from zero, $\Delta d = 0.16$, 95% CI [−0.13, 0.45].

Discussion

Whereas Studies 1 and 2 investigated the effects of upward and downward dimensional comparisons across different samples using an achievement situation in school, Study 3 replicated their findings in a generic achievement situation where participants received no further contextual information except a person's interval-scaled achievement in two unspecified achievement tests. The results of Study 3 confirmed the findings of Studies 1 and 2: Upward dimensional comparisons led to significantly lower other-rated self-concept, whereas downward dimensional comparisons led to significantly higher other-rated self-concept. The net effect of both comparison processes was again almost small positive, but not significantly different from zero, indicating both comparison directions to yield effects of nearly similar size.

There is one possible limitation to our results as the test scores used in Study 3 for both domains in the horizontal condition (*51* vs. *54*) were not perceived to be completely identical. However, our manipulation check confirmed that test scores in the standard domain were perceived as different achievement levels with the horizontal condition being located roughly in the middle between the other two. Therefore, its validity as a control group should not have been affected.

Study 4

To test whether the effects of upward and downward dimensional comparisons found in Studies 1 to 3 apply only to other-ratings or to students' actual (self-assessed) academic self-concepts in the same way (see the *Discussion* section of Study 1), we conducted a fourth study analyzing correlational data of a large German longitudinal study. Similar to the design of previous field studies by Pohlmann and Möller (2009) and to ensure comparability with the results of Studies 1 to 3, we only included students in our analysis that had received an average grade (3, "*satisfactory*") in German/mathematics and then assigned them to different achievement levels based on their grades in the other domain (below 3: *above-average*, 3: *average*, above 3: *below-average*). Like in Studies 1 to 3, this allowed us to examine the association between upward and downward dimensional comparisons and German and mathematical self-concepts separately while controlling for social comparison.

In addition, as the data in Study 4 were correlational in nature and the sample size was much larger compared with our experimental Studies 1 to 3, we were able to estimate the I/E model (Marsh, 1986). Because this path-analytical examination of dimensional comparisons is much more common than (quasi-) experimental studies (see Möller et al., 2009), estimating the I/E model is an important test whether the typical I/E effect pattern also replicates in our dataset. In the I/E model, contrasting effects of dimensional comparisons between two domains show in negative path coefficients of grades in one domain on self-concept in the other.

Method

Sample. The sample was taken from the German BiKS-8–14 study (e.g., Lorenz, Schmitt, Lehrl, Mudiappa, & Rossbach, 2013), a longitudinal study that aims at investigating students' academic development and school decisions from primary to secondary school.[4] Prior to their participation, informed consent was obtained from the students' parents. While the original BiKS-8–14 study included longitudinal data from $N = 3{,}288$ students from 3rd to 9th grade, we only examined 6th grade students' data at the beginning of secondary school because we believe students need time to acclimatize following the transition from primary (4th grade) to secondary school (5th grade). After deleting $N = 1{,}020$ students who did not provide data in 6th grade, a total of $N = 2{,}268$ students remained in the sample. The percentage of missing data for each of the self-concept items was low

[4] The research was supported by the German Research Foundation (DFG) AR 301/9-1 and AR 301/9-3.

and ranging between 1.5% and 2.2%. With regard to the grades in German and mathematics, information was missing for 7.6% of the participants. Sensitivity analysis showed that students with missing values within the variable 'grade mathematics' had parents with a lower HISEI than students without missing values within this variable, $M_{miss} = 50.23$, $M_{notMiss} = 53.58$, $p = .042$, $d = 0.21$. Moreover, students with missing data within the variable 'grade German' were more likely to have a migration background (at least one parent born abroad or not) than students without missing data within this variable; No Missing: 16.6%; Missing: 24.3%; $\chi^2 = 4.19$, $p = .047$, Cramer-V = .046. However, both effect sizes were only small.

In sum, sensitivity analysis suggests a missing at random (MAR) mechanism. Accordingly, a multiple imputation strategy was chosen for missing data treatment. Missing data values were imputed using the software SPSS. Multiple imputation is considered the gold-standard for modern missing data treatment (Enders, 2010). The imputation model included all self-concept items and grades as well as gender, the highest family ISEI score, a dichotomous school type indicator (academic track vs. nonacademic track) as well as a dichotomous indicator of immigrant background based on the parents' country of birth. A total of 10 data sets were imputed. For model parameters, t values, sample sizes, as well as effect sizes, we report pooled values as provided by the software SPSS. For F values and degrees of freedom, we report value ranges over the 10 imputed data sets because it is yet unclear how these values should best be pooled (Enders, 2010).

All of our analyses were thus based on a sample of $N = 2{,}268$ students, 50.8% female, $M = 12.40$ years old, $SD = 0.44$, 57.3% academic track, for which missing values were imputed. This sample served a basis for distributing the students to the upward, downward, or horizontal comparison condition (see description of the procedure below). As a robustness check, we replicated all analyses using a sample of $N = 2{,}046$ students based on listwise deletion of cases with missing data on the grades or on more than one of the self-concept indicators obtaining similar results.

Measures.

Achievement. Participants' grades in German and mathematics were used as achievement measures. Grades were acquired from participants' most recent school reports as reported by their teachers. If this information was missing, we used the grades which were reported by the students' parents.

Self-concept. Academic self-concept in German and mathematics was measured by three items per subject, adapted from the German BIJU study (Baumert, Gruehn, Heyn, Köller, & Schnabel, 1997) and the Self-Description Questionnaire (SDQ-II; Marsh, 1992), German self-concept: Cronbach's alpha = .89; Mathematics self-concept: Cronbach's alpha = .94. The items were in German and read: "*German/mathematics lessons are easy for me*," "*I learn new things quickly in German/mathematics classes*," and "*I am good at German/mathematics*." Participants rated how well the items applied to themselves on a 5-point scale from "*not at all*" to "*very much*." Accordingly, a higher score indicated a higher self-concept.

Procedure. Following the approach of the field studies conducted by Pohlmann and Möller (2009, Studies 1 and 2), we analyzed the association between upward and downward dimensional comparisons and German and mathematics self-concept separately. To this end, we included only participants in our analyses that had received an average grade (3, "*satisfactory*") in the target domain, that is, either German, average $N = 1{,}076.5$, or mathematics, average $N = 843.6$. The other domain served as the standard domain, respectively. Participants were then assigned to different achievement levels based on their grades in the standard domain: Students with a grade better than 3 ("*satisfactory*") were considered to be above-average, students with a grade equal to 3 were considered to be average, and students with a grade worse than 3 were considered to be below-average. The distribution of participants to the different achievement levels for both domains is shown in Table 5.

Holding achievement constant in the target domain (e.g., German/mathematics) made it possible to assess the covariation of achievement level in the standard domain (e.g., mathematics/German) and self-concept in the target domain. More precisely, the group with average German achievement gave us the opportunity to investigate the covariation of mathematics achievement with German self-concept, whereas the group with the average mathematics grade provided insight into the covariation of German achievement with mathematical self-concept. As shown in Figure 1, participants in the above-average group were assumed to perform upward dimensional comparisons from the target domain (where achievement was average) to the standard domain (where achievement did vary), whereas participants in the average group were assumed to perform horizontal dimensional comparisons, and participants in the below-average group were assumed to perform downward dimensional comparisons.

Results

Means and standard deviations for self-assessed self-concepts by comparison condition are presented in Table 6.

Preliminary analyses. Two ANOVAs revealed a significant main effect of achievement level in German (above-average vs. average vs. below-average) on self-concept in German, $F(2, 2267) = 131.6–151.8$, $p < .001$, $\eta^2 = .11$, and of achievement level in mathematics on self-concept in mathematics, $F(2, 2267) = 288.0–309.8$, $p < .001$, $\eta^2 = .21$. Post hoc Scheffé tests revealed that high achievement led to higher self-concept than average achievement, German: $M_{dif} = 0.49$, $p < .001$, $d = 0.61$, 95% CI [0.51, 0.70]; Mathematics: $M_{dif} = 0.69$, $p < .001$, $d = 0.73$, 95% CI [0.63, 0.83]. Low achievement led to lower self-concept than average achievement, German: $M_{dif} = -0.28$, $p < .001$, $d = -0.34$, 95% CI [−0.23, −0.45]; Mathematics: $M_{dif} = -0.56$, $p < .001$, $d = -0.55$, 95% CI [−0.45, −0.66].

To test whether an average grade (3, "*satisfactory*") led to similar self-concepts in both German and mathematics, a paired t test was conducted and showed self-concept in both domains to

Table 5
Study 4 Distribution of Participants to Achievement Levels

		Achievement in standard domain					
		Above-average		Average		Below-average	
Target domain	N	n	%	n	%	n	%
German	1,076.5	334.3	31.1	448.2	41.6	294.0	27.3
Mathematics	843.6	255.2	30.3	448.2	53.1	140.2	16.6

Note. If the subject in the target domain was German, the subject in the standard domain was mathematics and vice versa. Also, analyses were limited to participants with average achievement (3, "*satisfactory*") in the target domain. The sample sizes are averaged over the 10 imputed datasets.

Table 6
Study 4 Means and Standard Deviations of Self-Concept in the Target Domain by Dimensional Comparison Direction (Upward vs. Horizontal vs. Downward) and Subject in the Target Domain (German vs. Mathematics)

	Target domain			
	German		Mathematics	
Direction	*M* (*SD*)	*n*	*M* (*SD*)	*n*
Upward	3.21 (0.82)	334.3	3.06 (0.94)	255.2
Horizontal	3.42 (0.84)	448.2	3.28 (1.02)	448.2
Downward	3.61 (0.75)	294.0	3.57 (0.89)	140.2

Note. N = 2,268. If the subject in the target domain was German, the subject in the standard domain was mathematics and vice versa. Also, analyses were limited to participants with average achievement (3, "*satisfactory*") in the target domain.

differ significantly, $M_{dif} = -0.14$, $t(439-458) = -2.51$, $p = .012$, $d = -0.15$, 95% CI $[-0.02, -0.28]$. Accordingly, participants with an average achievement in both German and mathematics had a significantly better self-concept in German than in mathematics.

Pairwise comparisons. The same pairwise comparisons were conducted as described in Studies 1 to 3 and computed separately for German and mathematics self-concepts.

Upward dimensional comparisons. Participants reported significantly lower German/mathematical self-concept when achievement in mathematics/German was above-average than when it was average, German self-concept: $M_{dif} = -0.21$, $t(1066-1079) = -3.50$, $p < .001$, $d_{up} = -0.25$, 95% CI $[-0.39, -0.11]$; Mathematical self-concept: $M_{dif} = -0.22$, $t(832-848) = -2.71$, $p = .007$, $d_{up} = -0.22$, 95% CI $[-0.38, -0.07]$.

Downward dimensional comparisons. Participants also reported significantly higher German/mathematical self-concept when achievement in mathematics/German was below-average than when it was average, German self-concept: $M_{dif} = 0.19$, $t(1066-1079) = 3.02$, $p = .003$, $d_{down} = 0.24$, 95% CI [0.09, 0.38]; Mathematical self-concept: $M_{dif} = 0.29$, $t(832-848) = 3.08$, $p = .002$, $d_{down} = 0.29$, 95% CI [0.10, 0.48].

Net effect. Finally, the absolute effect sizes of upward, $d_{up} = -0.25$, and downward dimensional comparisons, $d_{down} = 0.24$, on German self-concept were not significantly different in size, $t(1066-1079) = 0.19$, $p = .850$, and their net effect was not different from zero, $\Delta d = -0.01$, 95% CI $[-0.15, 0.13]$. Similarly, the net effect of upward, $d_{up} = -0.22$, and downward dimensional comparisons, $d_{down} = 0.29$, on mathematical self-concept was not significantly different from zero, $\Delta d = 0.07$, 95% CI $[-0.10, 0.24]$, $t(1854) = 0.55$, $p = .585$.

I/E model. To test whether our data fit the I/E model (cf., Marsh, 1986; Möller et al., 2009), we conducted path analyses using structural equation modeling using M*plus* 7 (Muthèn & Muthèn, 2012). To handle missing data we used the Full Information Maximum Likelihood (FIML) procedure. This model-based approach to handling missing data is unbiased under the missing at random (MAR) assumption and retains statistical power as no observations are deleted. Because of these advantages, FIML is considered superior to traditional missing data treatment methods such as listwise deletion (Enders, 2010). As predicted by the I/E model, social comparisons led to positive standardized path coefficients from mathematical achievement to mathematical self-concept, $\beta = 0.55$, 95% CI [0.51, 0.60], $p < .001$, and from German achievement to German self-concept, $\beta = 0.43$, 95% CI [0.37, 0.48], $p < .001$. In addition, dimensional comparisons led to negative standardized path coefficients from German achievement to mathematical self-concept, $\beta = -0.16$, 95% CI $[-0.20, -0.12]$, $p < .001$, and from mathematical achievement to German self-concept, $\beta = -0.17$, 95% CI $[-0.22, -0.12]$, $p < .001$. Whereas the grades in German and mathematics were highly correlated, $r = .35$, 95% CI [0.29, 0.41], $p < .001$, the correlation of the self-concepts in both subjects was only small, $r = .15$, 95% CI [0.08, 0.22], $p < .001$. The model fit the data well, *CFI* = .995, *TLI* = .991, *RMSEA* = .032. Accordingly, our data fully supported the I/E model.

Discussion

Study 4 replicated the findings from our first three experimental studies by examining students' self-assessed self-concepts in school: Participants reported significantly lower self-concepts in the target domain when achievement in the standard domain was above-average (upward dimensional comparison) and significantly higher self-concepts when achievement in the standard domain was below-average (downward dimensional comparison). The results found were the same for the association of German achievement with mathematical self-concept and of mathematics achievement with German self-concept. Furthermore, the effect sizes of upward and downward dimensional comparisons were again not significantly different in size. Accordingly, their net effect was, though marginally positive, again near-zero. In addition, data in Study 4 supported the classic I/E model.

Whereas previous field studies by Pohlmann and Möller (2009, Studies 1 and 2) found only significant effects for downward dimensional comparisons, Study 4 is the first field study to show *both* upward and downward dimensional comparisons to be significantly related to students' self-assessed self-concepts in German and mathematics. Furthermore, Study 4 extends Studies 1 to 3 in terms of ecological validity and suggests that self- and other-ratings of self-concept yield similar results and that the cognitive processes underlying dimensional comparisons work similarly, regardless of the type of assessment made.

As a cross-sectional study, the causal interpretation of Study 4's correlational results obviously require caution. But even so, given the variety of experimental as well as longitudinal studies that have already provided causally interpretable evidence on the effects of dimensional comparisons in the past (e.g., Möller & Köller, 2001; Möller et al., 2011, 2014; Parker, Marsh, Morin, Seaton, & Van Zanden, 2015; Pohlmann & Möller, 2009; Strickhouser & Zell, 2015), we believe it rather safe to interpret the results in Study 4 as the effects of dimensional comparisons.

Because average achievement in German and mathematics led to different self-concepts, the validity of the horizontal comparison control group in Study 4 was again in question. But similar to the effects of achievement in the standard domain on self-concept in the standard domain in Study 3, the above-average, average, and below-average achievement levels in German/mathematics led to different verbal/mathematical self-concepts, with self-concepts of average performing students posited in the middle between the self-concepts of above-average and below-average performing students.

Study 5

In Study 5, we aimed at replicating Study 4's findings by using a larger sample and a different age group. To this end, we analyzed data of a second, representative German large-scale assessment study and split participants into groups by achievement level following the method detailed in Study 4. Again, we used the same design as shown in Figure 1.

As in Study 4, we also estimated the I/E model. Furthermore, as the sample size in Study 5 surpassed that of Study 4 and all achievement data was continuous (i.e., five grade levels), we were also able to look at (a) the increase or decrease in the effect sizes of upward and downward dimensional comparisons as a function of grade difference and (b) whether dimensional comparison effects could be shown not only for students with average grades in the target domain, but also with above or below average grades in the target domain. These analyses could thus provide additional insights on the generalizability of the effect patterns found.

Method

Sample. The sample was taken from the German national assessment study (IQB Ländervergleich; Pant et al., 2014) that was designed to compare ninth-grade students' academic achievement between the federal states of Germany based on national educational standards. In total, N = 44,584 9th grade students from 1,326 secondary schools took part in the study. The sample was representative for all German federal states as well as Germany as a whole. Informed consent was given by their parents prior to data collection. Because of a multiple matrix sampling design with different questionnaire versions, each participant completed only part of the item pool. Only N = 20,662 participants received questionnaire versions including self-concept measures. Of the N = 20,662 participants, 636 (3.1%) did not complete a single self-concept item. Sensitivity analysis showed that these students, on average, had slightly worse grades in mathematics, M_{miss} = 3.18, $M_{notMiss}$ = 2.98, $p < .01$, d = 0.13, and German, M_{miss} = 3.27, $M_{notMiss}$ = 3.14, $p < .01$, d = 0.24, and were more likely to be male (59.7%). Because these differences suggest a missing at random (MAR) mechanism, a multiple imputation strategy was chosen for missing data treatment as in Study 4. Missing data values were imputed using the R package *mice* (multiple imputation by chained equations; Van Buuren & Groothuis-Oudshoorn, 2011). The imputation model as well as the pooling procedure (using the software SPSS) were analogous to Study 4.

All our analyses are thus based on the sample of N = 20,662 students, 49.32% female, M = 15.05 years old, SD = 0.67; academic track: 37.76%, who received a self-concept questionnaire; for this sample all missing values were imputed. This sample served a basis for distributing the students to the upward, downward, or horizontal comparison condition. As a robustness check, we replicated all analyses using a sample of N = 18,635 students based on listwise deletion of cases with missing data on the grades or on more than one of the four self-concept indicators obtaining similar results.

Measures.

Achievement. Participants' grades in German and mathematics were acquired from their last end-of-term school reports as reported by school officials.

Self-concept. Academic self-concept in German and mathematics was measured by four items adapted from the international student questionnaire of the PISA study (Programme for International Student Assessment; OECD, 2003), German self-concept: Cronbach's alpha = .85; Mathematics self-concept: Cronbach's alpha = .90. Participants rated how well the items applied to themselves on a 4-point scale from 1 ("*strongly disagree*") to 4 ("*strongly agree*"). The items were in German and read: "*I am just not good at German/mathematics*," "*I get good marks in German/mathematics*," "*I learn German/mathematics quickly*," and "*I have always believed that German/mathematics is one of my best subjects*." The first item was inverse coded so that higher scores indicate higher self-concept.

Procedure. Identical to the approach used in Study 4 (with the exception of one additional analysis described below), we included only participants with an average grade (3, "*satisfactory*") in one of the subjects in our analyses, German: average N = 9,263.4; Mathematics: average N = 7,215.1. Participants were then assigned to different achievement levels based on their grades in the other subject. The distribution of participants to the different achievement levels for both subjects is shown in Table 7.

Results

Means and standard deviations for self-assessed self-concepts by comparison condition are presented in Table 8.

Preliminary analyses. Two ANOVAs revealed a significant main effect of achievement level in German (above-average vs. average vs. below-average) on self-concept in German, F(2, 20659) = 2,649.1–2,710.9, $p < .001$, $\eta^2 = .21$, and of achievement level in mathematics on self-concept in mathematics, F(2, 20659) = 4,630.2–4764.0, $p < .001$, $\eta^2 = .31$. Post hoc Scheffé tests revealed that high achievement led to higher self-concept than average achievement, German: M_{dif} = 0.46, $p < .001$, d = 0.77, 95% CI [0.74, 0.81]; Mathematics: M_{dif} = 0.65, $p < .001$, d = 0.92, 95% CI [0.89, 0.96]. Low achievement led to lower self-concept than average achievement, German: $M_{dif} = -0.36$, $p < .001$, $d = -0.59$, 95% CI [−0.62, −0.56]; Mathematics: $M_{dif} = -0.55$, $p < .001$, $d = -0.76$, CI 95% [−0.79, −0.73].

A paired t test was conducted to examine whether an average grade (3, "*satisfactory*") led to similar self-concepts in both German and mathematics. Results showed self-concept in both domains to not differ significantly, M_{dif} = 0.01, t(3637–3683) = 0.63, p = .529.

Table 7
Study 5 Distribution of Participants to Achievement Levels

		Achievement in standard domain					
		Above-average		Average		Below-average	
Target domain	N	n	%	n	%	n	%
German	9,263.4	2,161.6	23.3	3,661.1	39.5	3,440.7	37.1
Mathematics	7,215.1	1,940.3	26.9	3,661.1	50.7	1,613.7	22.4

Note. If the subject in the target domain was German, the subject in the standard domain was mathematics and vice versa. Also, analyses were limited to participants with average achievement (3, "*satisfactory*") in the target domain. The sample sizes are averaged over the 10 imputed datasets.

Table 8
Study 5 Means and Standard Deviations of Self-Concept in the Target Domain by Dimensional Comparison Direction (Upward vs. Horizontal vs. Downward) and Subject in the Target Domain (German vs. Mathematics)

	Target domain			
	German		Mathematics	
Direction	*M (SD)*	*n*	*M (SD)*	*n*
Upward	2.53 (0.59)	2,161.6	2.38 (0.69)	1,940.3
Horizontal	2.70 (0.57)	3,661.1	2.69 (0.71)	3,661.1
Downward	2.92 (0.57)	3,440.7	2.94 (0.68)	1,613.7

Note. $N = 20{,}662$. If the subject in the target domain was German, the subject in the standard domain was mathematics and vice versa. Also, analyses were limited to participants with average achievement (3, "*satisfactory*") in the target domain.

Pairwise comparisons. The same pairwise comparisons were conducted as described in the previous studies and computed separately for German and mathematics self-concept.

Upward dimensional comparisons. Self-concept in German/mathematics was lower when achievement in mathematics/German was above-average than when it was average, German self-concept: $M_{dif} = -0.16$, $t(9233–9293) = -10.86$, $p < .001$, $d_{up} = -0.29$, 95% CI [−0.35, −0.24]; Mathematical self-concept: $M_{dif} = -0.31$. $t(7198–7228) = -14.80$, $p < .001$, $d_{up} = -0.44$, 95% CI [−0.50, −0.39].

Downward dimensional comparisons. Self-concept in German/mathematics was higher when achievement in mathematics/German was below-average than when it was average, German self-concept: $M_{dif} = 0.21$, $t(9233–9293) = 15.29$, $p < .001$, $d_{down} = 0.39$, 95% CI [0.34, 0.43]; Mathematical self-concept: $M_{dif} = 0.25$, $t(7198–7228) = 11.69$, $p < .001$, $d_{down} = 0.36$, 95% CI [0.30, 0.42].

Net effect. The absolute mean differences resulting from upward and downward dimensional comparisons on German/mathematical self-concept did not differ significantly, German: $t(9233–9293) = 1.43$, $p = .154$; Mathematics: $t(7198–7228) = 1.70$, $p = .089$. Whereas the net effect of dimensional comparisons on German self-concept was marginally positive, $\Delta d = 0.09$, 95% CI [0.04, 0.14], the net effect on mathematical self-concept was marginally negative, $\Delta d = -0.08$, 95% CI [−0.14, −0.03]. Yet, both were very small and corresponded to net mean differences that were not significantly different from zero.

I/E model. Similar to the procedure in Study 4, we estimated the I/E model with the software *Mplus* as an additional analysis and used the Full Information Maximum Likelihood (FIML) procedure for missing data treatment (Enders, 2010). As in Study 4, the I/E model fit the data well, *CFI* = .957, *TLI* = .930, *RMSEA* = .078. Social comparisons led to positive standardized path coefficients from mathematical achievement to mathematical self-concept, $\beta = 0.70$, 95% CI [0.69, 0.72], $p < .001$, and from German achievement to German self-concept, $\beta = 0.63$, 95% CI [0.61, 0.64], $p < .001$. In addition, dimensional comparisons led to negative standardized path coefficients from German achievement to mathematical self-concept, $\beta = -0.23$, 95% CI [−0.24, −0.21], $p < .001$, and from mathematical achievement to German self-concept, $\beta = -0.25$, 95% CI [−0.26, −0.23], $p < .001$. Whereas the grades in German and mathematics were highly correlated, $r = .44$, 95% CI [0.43, 0.46], $p < .001$, the correlation between the self-concepts in both subjects was small negative, $r = -.15$, 95% CI [−0.17, −0.13], $p < .001$.

Grade difference effect. We examined the increase or decrease in the effect sizes of upward and downward dimensional comparisons as a function of grade difference. Results showed that the negative effect of upward dimensional comparisons increased with growing grade differences between the target (grade of 3) and the standard in German (grade of 1 in the standard: $M_{dif} = -0.30$, $t(9230–9290) = -7.64$, $p < .001$, $d_{up} = -0.53$, 95% CI [−0.66, −0.40]; grade of 2 in the standard: $M_{dif} = -0.16$, $t(9230–9290) = -9.51$, $p < .001$, $d_{up} = -0.27$, 95% CI [−0.33, −0.22]) as well as in mathematics (grade of 1 in the standard: $M_{dif} = -0.48$, $t(7195–7220) = -7.09$, $p < .001$, $d_{up} = -0.67$, 95% CI [−0.85, −0.50]; grade of 2 in the standard: $M_{dif} = -0.29$, $t(7195–7220) = -13.81$, $p < .001$, $d_{up} = -0.42$, 95% CI [−0.47, −0.36]). The same was mostly true for the positive effect of downward dimensional comparisons in German (grade of 4 in the standard: $M_{dif} = 0.19$, $t(9230–9290) = 12.79$, $p < .001$, $d_{down} = 0.34$, 95% CI [0.29, 0.39]; grade of 5 in the standard: $M_{dif} = 0.31$, $t(9230–9290) = 11.77$, $p < .001$, $d_{down} = 0.54$, 95% CI [0.46, 0.63]; grade of 6 in the standard: $M_{dif} = 0.37$, $t(9230–9290) = 2.94$, $p = .004$, $d_{down} = 0.65$, 95% CI [0.25, 1.05]) as well as in mathematics (grade of 4 in the standard: $M_{dif} = 0.25$, $t(7195–7220) = 11.08$, $p < .001$, $d_{down} = 0.35$, 95% CI [0.29, 0.41]; grade of 5 in the standard: $M_{dif} = 0.36$, $t(7195–7220) = 4.79$, $p < .001$, $d_{down} = 0.51$, 95% CI [0.31, 0.72]; grade of 6 in the standard: $M_{dif} = 0.07$, $t(7195–7220) = 0.102$, $p = .919$, $d_{down} = 0.10$, 95% CI [−1.50, 1.70]).

Different achievement levels. To make the results of this study easily comparable to the results of Studies 1 to 3, we held achievement in the target domain constant at an average level in our main analyses. However, as achievement data in Study 5 was continuous, this enabled us to test our hypotheses on different achievement levels, as well. To this end, we computed the effect sizes of upward and downward dimensional comparisons as well as their net effect also when participants had high achievement (i.e., a grade of 2, "*good*") or when they had low achievement (i.e., a grade of 4, "*sufficient*") in the target domain. Results for both high and low achievement levels were in line with our main analyses: Participants with a worse grade in the standard domain (downward dimensional comparison) reported higher self-concept in the target domain (high achievement: German self-concept: $M_{dif} = 0.21$, $t(5098–5137) = 11.86$, $p < .001$, $d_{down} = 0.36$, 95% CI [0.30, 0.42]; Mathematical self-concept: $M_{dif} = 0.27$, $t(4636–4675) = 12.98$, $p < .001$, $d_{down} = 0.41$, 95% CI [0.35, 0.48]; low achievement: German self-concept: $M_{dif} = 0.18$, $t(5034–5071) = 7.37$, $p < .001$, $d_{down} = 0.31$, 95% CI [0.23, 0.39]; Mathematical self-concept: $M_{dif} = 0.34$, $t(5906–5958) = 6.38$, $p < .001$, $d_{down} = 0.46$, 95% CI [0.33, 0.60]), whereas participants with a better grade in the standard domain (upward dimensional comparison) reported lower self-concept (high achievement: German self-concept: $M_{dif} = -0.14$, $t(5098–5137) = -4.39$, $p < .001$, $d_{up} = -0.24$, 95% CI [−0.33, −0.14]; Mathematical self-concept: $M_{dif} = -0.12$, $t(4636–4675) = -2.78$, $p = .006$, $d_{up} = -0.18$, 95% CI [−0.30, −0.05]; low achievement: German self-concept: $M_{dif} = -0.23$, $t(5034–5071) = -11.70$, $p < .001$, $d_{up} = -0.39$, 95% CI [−0.45, −0.33]; Mathematical

self-concept: $M_{dif} = -0.32$, $t(5906–5958) = -16.19$, $p < .001$, $d_{up} = -0.46$, 95% CI [−0.52, −0.41]). In line with our main analyses, the effect sizes of upward and downward dimensional comparisons were mostly of equal size and their net effect not significantly different from zero on a significance level of $\alpha = .1$ (high achievement: German self-concept: $\Delta d = 0.13$, 95% CI [0.05, 0.20], $t(5098–5137) = 1.72$, $p = .088$; Mathematical self-concept: $\Delta d = 0.24$, 95% CI [0.15, 0.33], $t(4636–4675) = 2.78$, $p = .006$; low achievement: German self-concept: $\Delta d = -0.08$, 95% CI [−0.15, −0.01], $t(5034–5071) = 1.15$, $p = .253$; Mathematical self-concept: $\Delta d = 0.00$, 95% CI [−0.09, 0.10], $t(5906–5958) = 0.21$, $p = .834$). Thus, the general pattern of findings in our main analyses was also generalizable to different achievement levels.

Discussion

In Study 5, we were able to replicate Study 4's results on the association between upward and downward dimensional comparisons and self-assessed self-concepts in German and mathematics in a different sample and an older age group. Once again, participants reported lower self-concepts in the upward dimensional comparison group and higher self-concepts in the downward dimensional comparison group resulting in a near-zero net effect of dimensional comparisons. Moreover, due to the rather large sample size in Study 5, we were also able to examine whether the magnitude of self-concept reduction and enhancement was related to the magnitude of actual difference in achievement levels. We not only found the effect sizes of upward and downward dimensional comparisons to increase with a growing grade difference between the target and the standard, but also found similar effect sizes of upward and downward comparisons (and therefore a near-zero net effect) on different achievement levels (i.e., high, average, and low achievement in the target). Furthermore, data in Study 5 also supported the classic I/E model.

Study 5 substantiates and extends the findings of Study 4 as well as the central assumption of DCT that upward and downward dimensional comparisons are significantly related to students' self-assessments of their self-concepts. Whereas Study 4 investigated these relations in 6th graders, Study 5 examined a representative sample of 9th graders. Therefore, Study 5 also provides further evidence of the generalizability of dimensional comparisons across age groups and achievement levels (see also Möller et al., 2009). Furthermore, and in line with the logic of the I/E model, Study 5 is the first to show the effect sizes of upward and downward dimensional comparisons to increase as a function of grade difference between the target and the standard.

General Discussion

The present research compared the effects of upward and downward dimensional comparisons in three experimental studies and two field studies. In line with our hypotheses, we found lower self-assessed and other-rated self-concepts in the target domain following upward dimensional comparisons and higher self-concepts following downward dimensional comparisons across all five studies. But in contrast to our expectations, the absolute effects of upward and downward dimensional comparisons were always of about equal size and their net effect near-zero.

Theoretical Implications

Our findings significantly substantiate and extend DCT in several key areas as they provide a new angle on the negative path coefficients found in the I/E model between the verbal and mathematical domains (e.g., Möller et al., 2009) as well as on the purpose of dimensional comparisons in general.

First, our findings provide strong empirical support to the central assumption of DCT that dimensional comparisons are indeed a "double-edged sword" (Möller & Marsh, 2013, p. 546) not only leading to self-concept enhancement in the better-off domain, but to self-concept reduction in the worse-off domain. Whereas previous studies found *either* significant effects for self-concept reduction following upward dimensional comparisons *or* for self-concept enhancement following downward dimensional comparisons (cf., Möller & Köller, 2001; Pohlmann & Möller, 2009), our five studies using experimental as well as correlational designs are the first to report statistically significant effects for *both* upward and downward dimensional comparisons. Moreover, we were also able to show that the magnitude of reductions and enhancements in self-concept covaried with the size of the achievement differences between two compared subjects, as the I/E model would predict.

Second, our results extend DCT significantly as we were able to find significant effects of similar size for upward *and* downward dimensional comparisons. We therefore consider it safe to assume that the net effect found in our studies was indeed near-zero and that its actual size or direction was of no relevance (see also Strickhouser & Zell, 2015). A meta-analysis comprising all previous results (see Table 1) as well as our own results from Studies 1 to 5 was able to confirm this assumption. Using a mixed effects model, we found the average positive effect of downward dimensional comparisons, $d_{down} = 0.47$, $SE = 0.05$, $p < .001$, 95% CI [0.38, 0.57], to be not significantly larger than the average negative effect of upward dimensional comparisons, $d_{up} = -0.36$, $SE = 0.05$, $p < .001$, 95% CI [−0.46, −0.27], and their net effect in total to be not significantly different from zero, $\Delta d = 0.05$, $SE = 0.03$, $p = .159$, 95% CI [−0.02, 0.12].[5] Our results therefore clearly refute one of DCT's 10 main hypotheses (cf., Möller et al., 2015), namely that dimensional comparisons are in total beneficial to the self. Instead, the present research suggests that the self-concept reducing effect of upward dimensional comparisons and the self-concept enhancing effect of downward dimensional comparisons effectively balance each other out (for similar results for social comparisons, see Schurtz, Pfost, & Artelt, 2014). Therefore, it seems more reasonable to assume that dimensional comparisons

[5] Absolute effect sizes of upward and downward dimensional comparisons did not differ in size, $\beta = -0.24$, $p = .119$, $\tau = .01$. Q tests indicated the effects of upward dimensional comparisons to be homogeneous, $Q_w(15) = 13.45$, $p = .567$, and the effects of downward dimensional comparisons to be significantly heterogeneous, $Q_w(15) = 24.89$, $p = .051$. Additional moderation analyses showed that the net effect of dimensional comparisons was affected by type of self-concept measure (evaluative vs. affective), $\beta = 0.53$, $p = .002$, $\tau = .01$, indicating that the net effect was significantly more positive when affective self-concept measures were used. However, the net effect was not moderated by varying sample size, $\beta = -0.27$, $p = .137$, $\tau = .02$, the type of assessment (self-assessments vs. other-ratings), $\beta = 0.17$, $p = .469$, $\tau = .01$, and the research design used (correlational vs. experimental), $\beta = 0.23$, $p = .185$, $\tau = .01$.

are *in general* neither positive nor hazardous to the self, but hold the potential for both. As stated by DCT, dimensional comparisons may indeed serve not only a compensatory self-enhancement function (e.g., Baumeister & Jones, 1978; Sedikides & Gregg, 2008), but may be driven by a self-evaluative motivation as well (e.g., Tesser, Millar, & Moore, 1988; Trope, 1975, 1980).

Moreover, the equal-sized effects of upward and downward dimensional comparisons along with their near-zero net effect found in our five studies point to another probable motivation triggering the use of dimensional comparisons, namely self-differentiation (cf., Möller et al., 2015). In other words, dimensional comparisons may help students not only to cope with low achievement feedback in one domain (by enhancing self-concept in the better-off domain via downward dimensional comparisons), but to also learn from it by polarizing their strengths and weaknesses (namely by reducing self-concept in the worse-off domain via upward dimensional comparisons) and facilitating important academic decisions like course choices (e.g., Köller et al., 2000; Nagy et al., 2006, 2008), how much time and effort to put into domain-specific tasks (e.g., Eccles et al., 1983; Wigfield & Eccles, 2000), and consequently even what career to pursue (e.g., Parker et al., 2012).

Third, our findings were generalizable and stable across different sample populations and age groups (high-school students vs. university students), domains (German/mathematics vs. generic achievement domains), achievement measures (school grades vs. interval-scaled achievement indicators), achievement levels (high vs. average vs. low), research designs (experimental vs. correlational designs), as well as types of assessment (self-assessments vs. other-ratings). Therefore, our results add to the growing body of research indicating the cognitive processes underlying dimensional comparisons to be highly generalizable to a variety of settings and constructs (cf., Möller et al., 2015; Möller & Marsh, 2013; Möller et al., 2009, 2016), and highlight their significance in comparative judgmental processes as assumed by DCT.

Limitations and Future Research

Although we found similar results regardless of the research design used (experimental vs. correlational approach) or type of assessment made (other-ratings vs. self-assessments), we investigated the effects of upward and downward dimensional comparisons only in experiments using other-rated self-concepts and in field studies using self-assessed self-concepts. Additional experimental studies on self-assessed self-concepts would provide causally interpretable data on the associations between upward and downward dimensional comparisons and self-assessed self-concepts found in our correlational designs in Studies 4 and 5, whereas field studies on other-rated self-concepts would examine whether dimensional comparisons are used by raters when inferring other people's self-concepts in the field (e.g., students inferring the self-concepts of their peers or teachers inferring their student's self-concepts in school). Though the present research is unable to close this gap, recently presented findings might. For example, a recently published study by Strickhouser and Zell (2015) examined the net effect of dimensional comparisons on self-assessed self-concept. The authors used a similar design as the present five studies (see Figure 1) and found the effects of upward, $d_{up} = -0.71$, and downward dimensional comparisons, $d_{down} = 0.64$, to be of equal size and their net effect, though marginally negative, $\Delta d = -0.07$, to not be significantly different from zero. Therefore, the results of Strickhouser and Zell (2015) on self-assessed self-concepts support the results of our Studies 1 to 3 on other-rated self-concept. Furthermore, Müller-Kalthoff et al. (2015) conducted a path-analytical study to investigate the effect of dimensional comparisons on other-rated self-concepts by asking teachers to infer the self-concepts of their high-school students in German and mathematics. Teachers inferred lower self-concept in both subjects when their students' grade in the respectively other subject was better, German: $\beta = -.05$; Mathematics: $\beta = -.14$. Therefore, the results of Müller-Kalthoff et al. (2015) indicate that the evaluative processes underlying social as well as dimensional comparisons produce similar outcomes for other-ratings as typically found for self-ratings (cf., Möller et al., 2009).

An additional limitation of the present research is that it looked exclusively at contrasting effects between dissimilar subjects (far comparisons), whereas DCT predicts assimilative effects for comparisons between similar subjects (near comparisons; see Marsh et al., 2014). Several field studies on the I/E model between closely related school subjects found effects ranging from smaller negative contrastive effects to even positive assimilative effects between the compared subjects (e.g., Jansen, Schroeders, Lüdtke, & Marsh, 2015; Marsh et al., 2014; Marsh, Lüdtke, et al., 2015; Möller, Streblow, Pohlmann, & Köller, 2006; Rost, Sparfeldt, Dickhäuser, & Schilling, 2005). Möller et al. (2016; see also Möller, Streblow, & Pohlmann, 2006) assume that dimensional comparison of a target and a standard is influenced by both domains' perceived similarity (for similar assumptions on social comparisons, see Mussweiler, 2003): When two subjects are rather dissimilar (e.g., mathematics and English), dissimilarity assumptions drive the evaluative process leading to contrast (i.e., negative path coefficients between achievement in one subject and self-concept in the other). As our Studies 1 to 5 show, this may lead to a near-zero net effect of dimensional comparisons. However, when two subjects are rather similar (e.g., mathematics and physics), Möller et al. (2016) assume similarity assumptions to drive the evaluative process leading to assimilation, instead (i.e., positive path coefficients). However, it is yet unclear whether, because of assimilative processes, the net effect of dimensional comparisons is equally near-zero or rather positive for near-comparisons.

Another limitation is that these studies looked exclusively at dimensional comparison of only two subjects at the same time, whereas it is plausible to assume that dimensional comparisons occur between several subjects and domains simultaneously in everyday life. Furthermore, it is yet unclear whether the net effect of dimensional comparisons is the same whether only two or more subjects are considered. Therefore, future field as well as experimental studies should consider the effects of dimensional comparisons between more than two subjects at the same time. In a similar vein, it would also be interesting to compare the effect sizes of upward and downward dimensional comparisons as well as their net effect to the effects of social and temporal comparisons within the same research design and sample. This would allow researchers and theorists to consider all factors involved in the formation of self-concepts simultaneously and examine whether their effects are independent or interact with each other.

Last, though we believe this to be a major contribution of the present research, our results indicated that the net effect of upward

and downward dimensional comparisons was near-zero, which is in conflict with previous findings by Pohlmann and Möller (2009) and Möller and Köller (2001, Study 2) who found either a positive or negative net effect of dimensional comparisons, respectively. At the same time, our results are in line with Strickhouser and Zell (2015, Study 1) who found the net effect of dimensional comparisons to be near-zero. Please also note that (except for Strickhouser & Zell, 2015, Study 1) none of the previous studies tested the difference of the effects of upward and downward dimensional comparisons (and therefore, their net effect) on statistical significance. The question remains why results in previous studies differ from ours. This is especially intriguing, as all previous studies as well as our Studies 1 to 5 used a similar design that allowed the examination of the effect of a variation of achievement in a standard domain on self-concept in a target domain (where achievement was held constant). Furthermore, all studies varied only minimally in terms of their method. Therefore, we would like to encourage future research to replicate our findings in different settings and with different constructs to shed more light on the possible moderators of the effects of dimensional comparisons.

Conclusions

In summary, the present research provides valuable insight into the effects of upward and downward dimensional comparisons indicating both to be meaningful determinants of students' academic self-concepts. Furthermore, our five studies indicate these effects to be equally strong and their net effect to be balanced between self-concept reduction and enhancement processes. Therefore, the present research closes a critical gap in the literature on dimensional comparison as previous studies found only significant effects for either upward or downward dimensional comparisons, but not for both, and DCT assumed their effects to be positive in total. We believe that our results make a strong case for the core assumption of DCT that dimensional comparisons are more than compensatory self-enhancement, namely a "double-edged sword" with equal potential to reduce as well as enhance self-concept in a target domain.

As positive self-concepts play a vital role in students' social and emotional development (Kagen, Moore, & Bredekamp, 1995) and may be considered a major objective for effective schooling worldwide (Brookover & Lezotte, 1979), dimensional comparisons may play a controversial role in academic and educational contexts. As Möller et al. (2015) put it, dimensional comparisons "often lead to an over- or underestimation of own abilities" (p. 434) in the better-off or worse-off domain, respectively. However, such estimations (in form of academic self-concepts) impact on students' academic choices and careers rather fundamentally: Students with higher self-concepts in a particular domain are more motivated (Eccles et al., 1983; Wigfield & Eccles, 2000), perform higher (Marsh & Craven, 2006; Möller et al., 2011, 2014; Niepel et al., 2014; Retelsdorf et al., 2014; Valentine et al., 2004), and are more likely to choose courses and careers in said domain (e.g., Köller et al., 2000; Nagy et al., 2006, 2008). Accordingly, negatively biased self-concepts following upward dimensional comparisons may lead students to invest less time and effort in the worse-off domain even if their achievements are actually not that bad. In contrast, positively biased self-concepts following downward dimensional comparisons may sway students to choose a career path in the better-off domain prematurely despite the worse-off domain being a reasonable choice as well. A critical example (cf., Jansen et al., 2015) may be the gender gap found in STEM subjects (e.g., Eccles, 2007; Wang, Eccles, & Kenny, 2013) where excelling in one domain (e.g., verbal subjects for most of the girls) may lead to overly skeptical evaluations of students' own abilities in another domain (e.g., physics). Moreover, dimensional comparisons may also affect variables like academic interest (Schurtz, Pfost, Nagengast, et al., 2014), achievement-related mood states (Möller & Husemann, 2006), test anxiety (Schilling, Sparfeldt, & John, 2005), and achievement-related satisfaction (Pohlmann & Möller, 2009).

Finally, we believe that self-concept–influencing processes like dimensional comparisons (as well as other comparison processes like social and temporal comparisons) should be considered by teachers, school counselors, and parents alike to aid students in developing realistic self-concepts in school and in choosing a fulfilling academic career for themselves. However, we also believe that prior to that, more research is necessary to ascertain whether preventing dimensional comparisons has a positive or negative overall effect on the academic self-concepts of students or whether dimensional comparison knowledge can really be used by teachers to help students gain a better or more realistic insight into their abilities.

References

Arens, A. K., & Möller, J. (2016). Dimensional comparisons in students' perceptions of the learning environment. *Learning and Instruction, 42,* 22–30. http://dx.doi.org/10.1016/j.learninstruc.2015.11.001

Baumeister, R. F. (1982). Self-esteem, self-presentation, and future interaction: A dilemma of reputation. *Journal of Personality, 50,* 29–45. http://dx.doi.org/10.1111/j.1467-6494.1982.tb00743.x

Baumeister, R. F., & Jones, E. E. (1978). When self-presentation is constrained by the target's knowledge: Consistency and compensation. *Journal of Personality and Social Psychology, 36,* 608–618. http://dx.doi.org/10.1037/0022-3514.36.6.608

Baumert, J., Gruehn, S., Heyn, S., Köller, O., & Schnabel, K.-U. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU)* [Educational and psychosocial development in adolescence]. Dokumentation, Band 1. Skalen Längsschnitt I, Welle 1–4. Berlin, Germany: Max-Plank-Institut für Bildungsforschung, Forschungsbereich "Erziehungswissenschaften und Bildungssysteme."

Biernat, M., & Eidelman, S. (2007). Standards. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 308–333). New York, NY: Guilford Press.

Boney-McCoy, S., Gibbons, F. X., & Gerrard, M. (1999). Self-esteem, compensatory self-enhancement, and the consideration of health risk. *Personality and Social Psychology Bulletin, 25,* 954–965. http://dx.doi.org/10.1177/01461672992511004

Brookover, W. B., & Lezotte, L. W. (1979). *Changes in school characteristics coincident with changes in student achievement.* East Lansing, MI: Institute for Research on Teaching.

Brown, J. D., & Smart, S. (1991). The self and social conduct: Linking self-representations to prosocial behavior. *Journal of Personality and Social Psychology, 60,* 368–375. http://dx.doi.org/10.1037/0022-3514.60.3.368

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dickhäuser, O. (2005). Teachers' inferences about students' self-concepts - The role of dimensional comparison. *Learning and Instruction, 15,* 225–235. http://dx.doi.org/10.1016/j.learninstruc.2005.04.004

Dickhäuser, O., Reuter, M., & Hilling, C. (2005). Coursework selection: A frame of reference approach using structural equation modelling. *British Journal of Educational Psychology, 75,* 673–688. http://dx.doi.org/10.1348/000709905X37181

Dickhäuser, O., Seidler, A., & Kölzer, M. (2005). Kein Mensch kann alles? Effekte dimensionaler Vergleiche auf das Fähigkeitsselbstkonzept [Nobody is perfect? Effects of dimensional comparisons on task-specific self-concepts]. *Zeitschrift für Pädagogische Psychologie / German Journal of Educational Psychology, 19,* 97–106. http://dx.doi.org/10.1024/1010-0652.19.12.97

Dodgson, P. G., & Wood, J. V. (1998). Self-esteem and the cognitive accessibility of strengths and weaknesses after failure. *Journal of Personality and Social Psychology, 75,* 178–197. http://dx.doi.org/10.1037/0022-3514.75.1.178

Eccles, J. S. (2007). Where are all the women? Gender differences in participation in physical science and engineering. In S. J. Ceci & W. M. Williams (Eds.), *Why aren't more women in science? Top researchers debate the evidence* (pp. 199–210). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/11546-016

Eccles, J. S., Adler, T. E., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75–146). San Francisco, CA: Freeman.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Epley, N., & Caruso, E. M. (2009). Perspective taking: Misstepping into others' shoes *Handbook of imagination and mental simulation* (pp. 295–309). New York, NY: Psychology Press. http://dx.doi.org/10.4135/9781412958479.n397

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87,* 327–339. http://dx.doi.org/10.1037/0022-3514.87.3.327

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140. http://dx.doi.org/10.1177/001872675400700202

Goetz, T., Frenzel, A. C., Hall, N. C., & Pekrun, R. (2008). Antecedents of academic emotions: Testing the internal/external frame of reference model for academic enjoyment. *Contemporary Educational Psychology, 33,* 9–33. http://dx.doi.org/10.1016/j.cedpsych.2006.12.002

Hirschmüller, S., Egloff, B., Nestler, S., & Back, M. D. (2013). The dual lens model: A comprehensive framework for understanding self-other agreement of personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 104,* 335–353. http://dx.doi.org/10.1037/a0030383

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

Jansen, M., Schroeders, U., Lüdtke, O., & Marsh, H. W. (2015). Contrast and assimilation effects of dimensional comparisons in five subjects: An extension of the I/E model. *Journal of Educational Psychology, 107,* 1086–1101. http://dx.doi.org/10.1037/edu0000021

Kagen, S. L., Moore, E., & Bredekamp, S. (1995). *Considering children's early development and learning: Toward common views and vocabulary.* Washington, DC: National Education Goals Panel.

Kirk, R. E. (2013). *Experimental design. Procedures for the behavioral sciences* (4th ed.). Thousand Oaks: Sage. http://dx.doi.org/10.4135/9781483384733

Köller, O., Daniels, Z., Schnabel, K. U., & Baumert, J. (2000). Kurswahlen von Mädchen und Jungen im Fach Mathematik: Zur Rolle von fachspezifischem Selbstkonzept und Interesse [Course selection of girls and boys in mathematics: The role of academic self-concept and interest]. *Zeitschrift für Pädagogische Psychologie / German Journal of Educational Psychology, 14,* 26–37. http://dx.doi.org/10.1024//1010-0652.14.1.26

Linville, P. W. (1985). Self-complexity and affective extremity: Don't put all of your eggs in one cognitive basket. *Social Cognition, 3,* 94–120. http://dx.doi.org/10.1521/soco.1985.3.1.94

Lorenz, C., Schmitt, M., Lehrl, S., Mudiappa, M., & Rossbach, H.-G. (2013). The Bamberg BiKS research group. In M. Pfost, C. Artelt, & S. Weinert (Eds.), *The development of reading literacy from early childhood to adolescence. Empirical findings from the Bamberg BiKS longitudinal studies* (pp. 15–34). Bamberg, Germany: University of Bamberg Press.

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23,* 129–149. http://dx.doi.org/10.3102/00028312023001129

Marsh, H. W. (1992). *Self-Description Questionnaire (SDQ) II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept. A test manual and research monograph.* Sydney, Australia: University of Western Sydney Macarthur, Faculty of Education.

Marsh, H. W. (2006). *Self-concept theory, measurement and research into practice: The role of self-concept in Educational Psychology*. Leicester, UK: British Psychological Society.

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., Nagengast, B., . . . Abu-Hilal, M. M. (2015). The internal/external frame of reference model of self-concept and achievement relations: Age-cohort and cross-cultural differences. *American Educational Research Journal, 52,* 168–202. http://dx.doi.org/10.3102/0002831214549453

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163. http://dx.doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Kuyper, H., Seaton, M., Parker, P. D., Morin, A. J. S., Möller, J., & Abduljabbar, A. S. (2014). Dimensional comparison theory: An extension of the internal/external frame of reference effect on academic self-concept formation. *Contemporary Educational Psychology, 39,* 326–341. http://dx.doi.org/10.1016/j.cedpsych.2014.08.003

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Abduljabbar, A. S., Abdelfattah, F., & Jansen, M. (2015). Dimensional comparison theory: Paradoxical relations between self-beliefs and achievements in multiple domains. *Learning and Instruction, 35,* 16–32. http://dx.doi.org/10.1016/j.learninstruc.2014.08.005

Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20,* 107–123. http://dx.doi.org/10.1207/s15326985ep2003_1

Möller, J. (1999). Soziale, fachbezogene und temporale Vergleichsprozesse bei der Beurteilung schulischer Leistung [Social, subject-specific, and temporal comparison processes in the evaluation of academic achievements]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 31,* 11–17. http://dx.doi.org/10.1026//0049-8637.31.1.11

Möller, J. (2005). Paradoxical effects of praise and criticism: Social, dimensional and temporal comparisons. *British Journal of Educational Psychology, 75,* 275–295. http://dx.doi.org/10.1348/000709904X24744

Möller, J., Helm, F., Müller-Kalthoff, H., Nagy, N., & Marsh, H. W. (2015). Dimensional comparisons and their consequences for self-concept, motivation, and emotion. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 430–436). Oxford, UK: Elsevier. http://dx.doi.org/10.1016/B978-0-08-097086-8.26092-3

Möller, J., & Husemann, N. (2006). Internal comparisons in everyday life. *Journal of Educational Psychology, 98,* 342–353. http://dx.doi.org/10.1037/0022-0663.98.2.342

Möller, J., & Köller, O. (1997). Kontexteffekte in Berichtszeugnissen [Context effects in qualitative student records]. *Psychologie in Erziehung und Unterricht, 44,* 187–196.

Möller, J., & Köller, O. (2001). Dimensional comparisons: An experimental approach to the internal/external frame of reference model. *Journal of Educational Psychology, 93,* 826–835. http://dx.doi.org/10.1037/0022-0663.93.4.826

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review, 120,* 544–560. http://dx.doi.org/10.1037/a0032459

Möller, J., Müller-Kalthoff, H., Helm, F., Nagy, N., & Marsh, H. W. (2016). The generalized internal/external frame of reference model: An extension to dimensional comparison theory. *Frontline Learning Research, 4,* 1–11.

Möller, J., & Pohlmann, B. (2010). Achievement differences and self-concept differences: Stronger associations for above or below average students? *British Journal of Educational Psychology, 80,* 435–450. http://dx.doi.org/10.1348/000709909X485234

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A Meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79,* 1129–1167. http://dx.doi.org/10.3102/0034654309337522

Möller, J., Retelsdorf, J., Köller, O., & Marsh, H. W. (2011). The reciprocal internal/external frame of reference model: An integration of models of relations between academic achievement and self-concept. *American Educational Research Journal, 48,* 1315–1346. http://dx.doi.org/10.3102/0002831211419649

Möller, J., & Savyon, K. (2003). Not very smart, thus moral: Dimensional comparisons between academic self-concept and honesty. *Social Psychology of Education, 6,* 95–106. http://dx.doi.org/10.1023/A:1023247910033

Möller, J., Streblow, L., & Pohlmann, B. (2006). The belief in a negative interdependence of math and verbal abilities as determinant of academic self-concepts. *British Journal of Educational Psychology, 76,* 57–70. http://dx.doi.org/10.1348/000709905X37451

Möller, J., Streblow, L., Pohlmann, B., & Köller, O. (2006). An extension to the internal/external frame of reference model to two verbal and numerical domains. *European Journal of Psychology of Education, 21,* 467–487. http://dx.doi.org/10.1007/BF03173515

Möller, J., Zimmermann, F., & Köller, O. (2014). The reciprocal internal/external frame of reference model using grades and test scores. *British Journal of Educational Psychology, 84,* 591–611. http://dx.doi.org/10.1111/bjep.12047

Müller-Kalthoff, H., Helm, F., & Möller, J. (2015, August 21st). *When knowledge is power: On dimensional comparison processes in teachers' inferences of students' self-concepts.* Paper presented at the SELF biennial international conference, Kiel, Germany.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review, 110,* 472–489. http://dx.doi.org/10.1037/0033-295X.110.3.472

Muthèn, L. K., & Muthèn, B. O. (2012). *Mplus. Statistical analysis with latent variables. User's guide* (7th ed.). Los Angeles, CA: Author.

Nagy, G., Garrett, J., Trautwein, U., Cortina, K. S., Baumert, J., & Eccles, J. S. (2008). Gendered high school course selection as a precursor of gendered careers: The mediating role of self-concept and intrinsic value. In H. M. G. Watt & J. S. Eccles (Eds.), *Gender and occupational outcomes. Longitudinal assessments of individual, social, and cultural influences* (pp. 115–143). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/11706-004

Nagy, G., Trautwein, U., Baumert, J., Köller, O., & Garrett, J. (2006). Gender and course selection in upper secondary education: Effects of academic self-concept and intrinsic value. *Educational Research and Evaluation, 12,* 323–345. http://dx.doi.org/10.1080/13803610600765687

Niepel, C., Brunner, M., & Preckel, F. (2014). The longitudinal interplay of students' academic self-concepts and achievements within and across domains: Replicating and extending the reciprocal internal/external frame of reference model. *Journal of Educational Psychology, 106,* 1170–1191. http://dx.doi.org/10.1037/a0036307

OECD. (2003). *PISA 2003 Tech.* Rep. No. Retrieved from http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/35188570.pdf

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2014). *The IQB national assessment study 2012. Competencies in mathematics and the sciences at the end of secondary level I.* Retrieved from https://www.iqb.hu-berlin.de/laendervergleich/laendervergleich/lv2012/Bericht/IQB_NationalAsse.pdf

Parker, P. D., Marsh, H. W., Morin, A. J., Seaton, M., & Van Zanden, B. (2015). If one goes up the other must come down: Examining ipsative relationships between math and English self-concept trajectories across high school. *British Journal of Educational Psychology, 85,* 172–191. http://dx.doi.org/10.1111/bjep.12050

Parker, P. D., Schoon, I., Tsai, Y.-M., Nagy, G., Trautwein, U., & Eccles, J. S. (2012). Achievement, agency, gender, and socioeconomic background as predictors of postschool choices: A multicontext study. *Developmental Psychology, 48,* 1629–1642. http://dx.doi.org/10.1037/a0029167

Pohlmann, B., & Möller, J. (2006). Vergleichseffekte auf kognitive, affektive und motivationale Variablen. Eine experimentelle Überprüfung [Comparison effects on cognitive, affective, and motivational variables: An experimental examination]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 38,* 79–87. http://dx.doi.org/10.1026/0049-8637.38.2.79

Pohlmann, B., & Möller, J. (2009). On the benefit of dimensional comparisons. *Journal of Educational Psychology, 101,* 248–258. http://dx.doi.org/10.1037/a0013151

Retelsdorf, J., Köller, O., & Möller, J. (2014). Reading achievement and reading self-concept—Testing the reciprocal effects model. *Learning and Instruction, 29,* 21–30. http://dx.doi.org/10.1016/j.learninstruc.2013.07.004

Rost, D. H., Sparfeldt, J. R., Dickhäuser, O., & Schilling, S. R. (2005). Dimensional comparisons in subject-specific academic self-concepts and achievements: A quasi-experimental approach. *Learning and Instruction, 15,* 557–570. http://dx.doi.org/10.1016/j.learninstruc.2005.08.003

Schilling, S. R., Sparfeldt, J. R., & John, M. (2005). Besser in Mathe - besorgter in Deutsch? Beziehungen zwischen Schulleistungen, schulischen Selbstkonzepten und Prüfungsängsten im Rahmen des I/E-Modells. [Better in math - more worried in German? Relationships between academic achievement, academic self-concepts, and test anxieties in the context of the I/E model] In S. R. Schilling, J. R. Sparfeldt, & C. Pruisken (Eds.), *Aktuelle Aspekte pädagogisch-psychologischer Forschung* (pp. 159–178). Münster, Germany: Waxmann.

Schurtz, I. M., Pfost, M., & Artelt, C. (2014). Variieren die Selbstkonzeptdifferenzen in Abhängigkeit vom Leistungsniveau? Differenzielle Zusammenhänge in Deutsch, Englisch und Mathematik [Do self-concept differences vary in dependence of the achievement level? Differential relations in German, English, and mathematics]. *Zeitschrift für Pädagogische Psychologie / German Journal of Educational Psychology, 28,* 31–42. http://dx.doi.org/10.1024/1010-0652/a000122

Schurtz, I. M., Pfost, M., Nagengast, B., & Artelt, C. (2014). Impact of social and dimensional comparisons on student's mathematical and English subject-interest at the beginning of secondary school. *Learning and Instruction, 34,* 32–41. http://dx.doi.org/10.1016/j.learninstruc.2014.08.001

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 3,* 102–116. http://dx.doi.org/10.1111/j.1745-6916.2008.00068.x

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46,* 407–441. http://dx.doi.org/10.3102/00346543046003407

Skaalvik, E. M., & Skaalvik, S. (2002). Internal and external frames of reference for academic self-concept. *Educational Psychologist, 37,* 233–244. http://dx.doi.org/10.1207/S15326985EP3704_3

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psycholgoy [Social psychology studies of the self: Perspectives and programs]* (Vol. 21, pp. 261–302). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/S0065-2601(08)60229-4

Strickhouser, J. E., & Zell, E. (2015). Self-evaluative effects of dimensional and social comparison. *Journal of Experimental Social Psychology, 59,* 60–66. http://dx.doi.org/10.1016/j.jesp.2015.03.001

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103,* 193–210. http://dx.doi.org/10.1037/0033-2909.103.2.193

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. *Advances in Experimental Social Psychology, 21,* 181–227. http://dx.doi.org/10.1016/S0065-2601(08)60227-0

Tesser, A., Millar, M., & Moore, J. (1988). Some affective consequences of social comparison and reflection processes: The pain and pleasure of being close. *Journal of Personality and Social Psychology, 54,* 49–61. http://dx.doi.org/10.1037/0022-3514.54.1.49

Tietjens, M., & Niewerth, J. (2005). Effekte sozialer und dimensionaler Vergleichsinformationen im Sport [Social and dimensional comparisons in a sport-related context]. *Zeitschrift für Sportpsychologie, 12,* 2–10. http://dx.doi.org/10.1026/1612-5010.12.1.2

Trope, Y. (1975). Seeking information about one's ability as a determinant of choice among tasks. *Journal of Personality and Social Psychology, 32,* 1004–1013. http://dx.doi.org/10.1037/0022-3514.32.6.1004

Trope, Y. (1980). Self-assessment, self-enhancement, and task preference. *Journal of Experimental Social Psychology, 16,* 116–129. http://dx.doi.org/10.1016/0022-1031(80)90003-7

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39,* 111–133. http://dx.doi.org/10.1207/s15326985ep3902_3

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). *MICE: Multivariate imputation by chained equations in R. 45,* 67. http://dx.doi.org/10.18637/jss.v045.i03

Wang, M.-T., Eccles, J. S., & Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science, 24,* 770–775. http://dx.doi.org/10.1177/0956797612458937

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25,* 68–81. http://dx.doi.org/10.1006/ceps.1999.1015

Received June 16, 2015
Revision received October 30, 2016
Accepted November 11, 2016 ■

Call for Papers
A Focused Collection of Qualitative Studies in the Psychological Sciences: Reasoning and Participation in Formal and Informal Learning Environments

Journal of Educational Psychology
Guest Editors: Tanner LeBaron Wallace and Eric Kuo

Reasoning and participation are two central topics of education research in the psychological sciences. Understanding the mechanisms that govern thought and reasoning has long been a core enterprise of educational psychology and, over time, more modern views on learning have promoted participation as a key feature for research—either as a facilitator of learning, a practice to be learned, or as an operationalization of learning itself.

We are pleased to announce a focused collection highlighting qualitative studies of reasoning and participation in formal and informal learning environments. By inviting studies incorporating qualitative methods, we aim to complement the experimental and longitudinal statistical research on these topics that is typically published in this journal. We encourage submission of papers focused on the following (or closely related) topics:

- Student reasoning and/or participation in novel learning environments or activities
- The relations between student reasoning, motivation, identity, and participation
- Student perceptions and meaning-making during participatory experiences
- Dynamic models of student reasoning that are grounded in data
- Explanatory accounts for how and why participation is successful (or not)
- Identifying new goals or targeted outcomes for reasoning or participation

We especially welcome qualitative studies that demonstrate the possibilities for unique discovery afforded by inductive analysis of rich data sources (e.g., real-time recordings of student reasoning, participation, discourse, and physical action, students' meaning-making anchored to particular interactions experienced). This collection will highlight the benefits of qualitative methods for extending and deepening theoretical and empirical understandings of reasoning and participation in both formal and informal learning environments.

The deadline for manuscript submissions is **March 1, 2018**. We invite authors to contact the Guest Editors of this collection, Tanner LeBaron Wallace (twallace@pitt.edu) and Eric Kuo (erickuo@pitt.edu), for discussion on how to maximize alignment between their submissions and this focused collection, though it is not required. Please follow both APA guidelines as well as specific submission criteria for the journal. When submitting manuscripts, please also indicate your intent to submit to this focused collection in the required cover letter.

All manuscripts must be submitted electronically at http://www.editorialmanager.com/edu. In the submission portal, please select the article type "Special Section: Reasoning & Participation – Qualitative." For more information on the *Journal of Educational Psychology*, please visit http://www.apa.org/pubs/journals/edu/.

New Editors Appointed, 2019–2024

The Publications and Communications Board of the American Psychological Association announces the appointment of 11 new editors for 6-year terms beginning in 2019. As of January 1, 2018, new manuscripts should be directed as follows:

- *Clinician's Research Digest: Adult Populations and Clinician's Research Digest: Child and Adolescent Populations* (http://www.apa.org/pubs/journals/crd/ and http://www.apa.org/pubs/journals/crp/), **Marisol Perez, PhD,** Arizona State University
- *Journal of Experimental Psychology: Learning, Memory, and Cognition* (http://www.apa.org/pubs/journals/xlm/), **Aaron S. Benjamin, PhD,** University of Illinois at Urbana-Champaign
- *Journal of Neuroscience, Psychology, and Economics* (http://www.apa.org/pubs/journals/npe/), **Samuel M. McClure, PhD,** Arizona State University
- *Journal of Threat Assessment and Management* (http://www.apa.org/pubs/journals/tam/), **Laura S. Guy, PhD,** Protect International Risk and Safety Services Inc., Vancouver, BC, Canada
- *Professional Psychology: Research and Practice* (http://www.apa.org/pubs/journals/pro/), **Kathi A. Borden, PhD,** Antioch University New England
- *Psychiatric Rehabilitation Journal* (http://www.apa.org/pubs/journals/prj/), **Sandra G. Resnick, PhD,** VA Connecticut Healthcare System
- *Psychology and Aging* (http://www.apa.org/pubs/journals/pag/), **Elizabeth A. L. Stine-Morrow, PhD,** University of Illinois at Urbana-Champaign
- *Psychology of Violence* (http://www.apa.org/pubs/journals/vio/), **Antonia Abbey, PhD,** Wayne State University
- *Psychology, Public Policy, and Law* (http://www.apa.org/pubs/journals/law/), **Michael E. Lamb, PhD,** University of Cambridge
- *Training and Education in Professional Psychology* (http://www.apa.org/pubs/journals/tep/), **Debora J. Bell, PhD,** University of Missouri-Columbia
- *Traumatology* (http://www.apa.org/pubs/journals/trm/), **Regardt J. Ferreira, PhD,** Tulane University

Current editors Thomas E. Joiner, PhD, Robert Greene, PhD, Daniel Houser, PhD, and Bernd Weber, PhD, Stephen D. Hart, PhD, Ronald T. Brown, PhD, Judith A. Cook, PhD, and Kim T. Mueser, PhD, Ulrich Mayr, PhD, Sherry Hamby, PhD, Michael E. Lamb, PhD, Michael C. Roberts, PhD, and Brian E. Bride, PhD, will receive and consider new manuscripts through December 31, 2017.

Instructions to Authors
Journal of Educational Psychology
www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/pubs/journals/edu. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychologial Bulletin, 139,* 133–151. http://dx.doi.org/10.1037/a0028566

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied in TIFF or EPS format. APA's policy on publication of color figures is available at http://www.apa.org/pubs/authors/instructions.aspx?item=6.

Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/pubs/authors/posting.aspx. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . .." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . .." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

Permissions. Authors of accepted papers must obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including test materials (or portions thereof), photographs, and other graphic images (including those used as stimuli in experiments). On advice of counsel, APA may decline to publish any image whose copyright status is unknown.

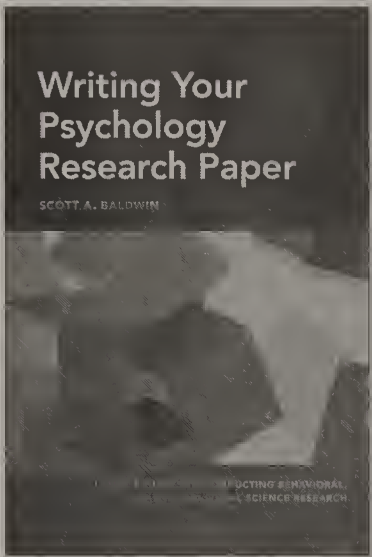
Supplemental materials. APA can place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/pubs/authors/supp-material.aspx for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/pubs/journals/edu/index.aspx (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the editorial office at CJohnson@apa.org.

AMERICAN PSYCHOLOGICAL ASSOCIATION

WRITING YOUR PSYCHOLOGY RESEARCH PAPER

Scott A. Baldwin

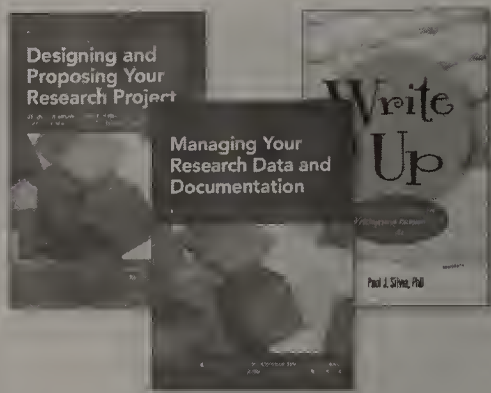


This encouraging primer for undergraduates explains how to write a clear, compelling, well-organized research paper. From picking a promising topic, to finding and digesting the pertinent literature, to developing a thesis, to outlining and presenting ideas, to editing for clarity and concision—each step is broken down and illustrated with examples. A bonus chapter discusses how to combat procrastination. Students learn that the best writing is done in chunks over long periods of time, and that writing is a skill that improves with practice. *2018. 126 pages. Paperback.* **Series: *Concise Guides to Conducting Behavioral, Health, and Social Science Research***

List: $29.95 | APA Member/Affiliate: $25.95 | ISBN 978-1-4338-2707-5 | Item # 4313044

CONTENTS

ALSO OF INTEREST



Designing and Proposing Your Research Project

Jennifer Brown Urban and Bradley Matheus Van Eeden-Moorefield

2018. 139 pages. Paperback.

Series: *Concise Guides to Conducting Behavioral, Health, and Social Science Research*

List: $29.95 | APA Member/Affiliate: $25.95
ISBN 978-1-4338-2708-2 | Item # 4313045

Managing Your Research Data and Documentation

Kathy R. Berenson

2018. 117 pages. Paperback.

Series: *Concise Guides to Conducting Behavioral, Health, and Social Science Research*

List: $29.95 | APA Member/Affiliate: $25.95
ISBN 978-1-4338-2709-9 | Item # 4313048
AVAILABLE ON AMAZON KINDLE®

AN APA LIFETOOLS® BOOK

Write It Up

Practical Strategies for Writing and Publishing Journal Articles

Paul J. Silvia, PhD

2015. 224 pages. Paperback.

List: $19.95 | APA Member/Affiliate: $19.95
ISBN 978-1-4338-1814-1 | Item # 4441024
AVAILABLE ON AMAZON KINDLE®

PsycBOOKS®
Access to chapters from a variety of APA scholarly & professional books.

APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books
In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502
In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD3178